AD-A008 850

DAMAGE PROFILES IN SILICON AND THEIR IMPACT ON DEVICE RELIABILITY

Guenter H. Schwuttke

International Business Machines Corporation

Prepared for:

Advanced Research Projects Agency

1 February 19/5

Unclassified

Security Classification

AD-A008 850

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*
International Business Machines Corporation
System Products Division, East Fishkill
Hopewell Junction, N.Y. 12533

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP

3. REPORT TITLE
DAMAGE PROFILES IN SILICON
AND THEIR IMPACT ON DEVICE RELIABILITY

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*
Scientific July 1974 to December 1974

5. AUTHOR(S) *(First name, middle initial, last name)*
Guenter H. Schwuttke

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
1 February 1975	63	19

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
DAHC 15-72-C-0274	TR 22.1865
b. PROJECT, TASK, WORK UNIT NOS.	
c. DOD ELEMENT	9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*
d. DOD SUBELEMENT	

10. DISTRIBUTION STATEMENT

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Advanced Research Projects Agency

13. ABSTRACT

This report consists of two parts. Part one describes a new approach to eliminate crystal defects in silicon wafers due to high temperature batch processing. It is shown that considerable improvements in crystal perfection are obtained if a closed boat is used for wafer processing. The second part describes an efficient technique that permits measurements of generation lifetime in silicon in the range of $\gtrsim$ 1msec to $<<$ 0.1nsec. The technique is suitable for process characterization.

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Silicon wafer						
Silicon wafer processing						
Silicon defect						
Surface state density						
Lifetime						

CONTENTS PAGE

LIST OF INVESTIGATORS

The project is supervised by Dr. G. H. Schwuttke, principal investigator. The following people contributed to the work in the report:

Dr. W. Fahrner	-	Investigator (Visiting Scientist)
Mr. E. W. Hearn	-	Investigator
Dr G. H. Schwuttke	-	Investigator
Dr. E. H. teKaat	-	Investigator (Visiting Scientist)
Mr. J. F. Francis	-	Technical Support
Mr. C. P. Schneider	-	Technical Support
Mr. H. L. Stellefson	-	Technical Support

SUMMARY

This report consists of two parts. Part one describes a new approach for semiconductor batch processing and is concerned with the reduction or elimination of crystal defects due to high temperature processing. Temperature gradient measurements and x-ray topography are used to characterize silicon wafers before and after batch processing. Experiments are performed at a temperature range from 900°C to 1200°C. Wafers are processed using standard and modified quartz boats. Temperature gradients in wafers during heat cycling are measured and related to crystal perfection. It is shown that defects in wafers due to heat cycling are substantially reduced through closed boat processing. Similar investigations are reported for wafer warpage. Finally, successful processing of large area silicon slabs (115mm x 57mm area and 0.375mm thick) is discussed.

The second part describes an efficient measurement technique that permits measurements of generation lifetime in the range of $\gtrsim$1msec to $\ll$0.1nsec. For the measurements, a metal oxide semiconductor (MOS) capacitor is biased into strong inversion and subsequently switched into deep depletion. An appropriate experimental setup prints out or displays a typical recovery time which is introduced into a computer fed by the theoretical generation model and the wafer data. The technique is suitable for process characterization.

Chapter 1

THE CLOSED BOAT:

A NEW APPROACH FOR SEMICONDUCTOR BATCH PROCESSING

by

E. W. Hearn, E. H. teKaat and G. H. Schwuttke

1. INTRODUCTION

The influence of process generated crystallographic defects in silicon wafers on device yield is well established (1,2). The importance of the batch concept for modern semiconductor manufacturing and the resulting degradation of silicon wafer perfection during batch processing has also been pointed out (3,4). Accordingly, high temperature processes, such as oxidations and diffusions lead to interesting temperature effects if a row of wafers is processed in a boat. Such effects can degrade the crystalline perfection of a silicon slice. The amount of degradation introduced into one slice depends on several parameters such as position of the slice in the row, distance between the single wafers in the row, number of wafers in the row and diameter of wafers in the row. Consequently, the amount of deformation per wafer can vary considerably.

This paper describes an improved method for heating and cooling semiconductor wafers. It is shown that the closed boat applied to semiconductor wafer processing minimizes or eliminates thermally induced defects in silicon wafers processed under manufacturing conditions(5).

2. EXPERIMENTAL PROCEDURE

57.2mm (2 1/4 inch) diameter silicon wafers were prepared mainly from [100] crystals grown by the Czochralski technique. The wafers were all chemical-mechanical polished to a thickness of approximately 300μm. Prior to any heat treatment the wafers were all dislocation free as revealed by x-ray topography. Experiments were conducted to investigate the influence of temperature gradients on wafer perfection. Such gradients are generated in silicon during rapid cooling from high temperature to room temperature. For the experiments batches of wafers were loaded into standard oxidation quartz boats. Each batch consisted of 100 wafers. A batch contained at least 15 specimen wafers positioned symmetrically in the center of the boat supplemented by 85 dummy wafers positioned around the specimen wafers. The wafers in the boat were spaced 0.2mm apart. The experiments were done in an oxidizing atmosphere

at temperatures of 900°C, 1100°C and 1200°C. The wafers were pushed into the hot zone of the furnace without any special precautions. Likewise the loaded boat was removed from the furnace after it had reached the furnace temperature by pulling it rapidly out to room temperature. Thermal gradients arising in wafers during the critical cooling period were measured by a differential thermocouple technique. Diagnostic x-ray charts were recorded through SOT topography (6) and used to assess crystallographic perfection of wafers. Similar measurements were made with closed boats. The closed boat consists in its simplest form of a standard boat covered by a quartz tube cut in half and turned upside down over the standard boat as a roof to close the standard boat. A standard boat is shown in the photograph of Fig. 1a. An example of a closed boat obtained from a standard boat by covering one half of it with a roof is shown in the picture of Fig. 1b. Such boats allow simultaneous processing of wafers in open and closed boat configuration. Another type of closed boat, also used for the experiments, is shown in Fig. 1c. and Fig. 1d.

3. RESULTS

Drastic radial temperature changes occur in silicon wafers stacked in a batch when quickly pulled out of a high

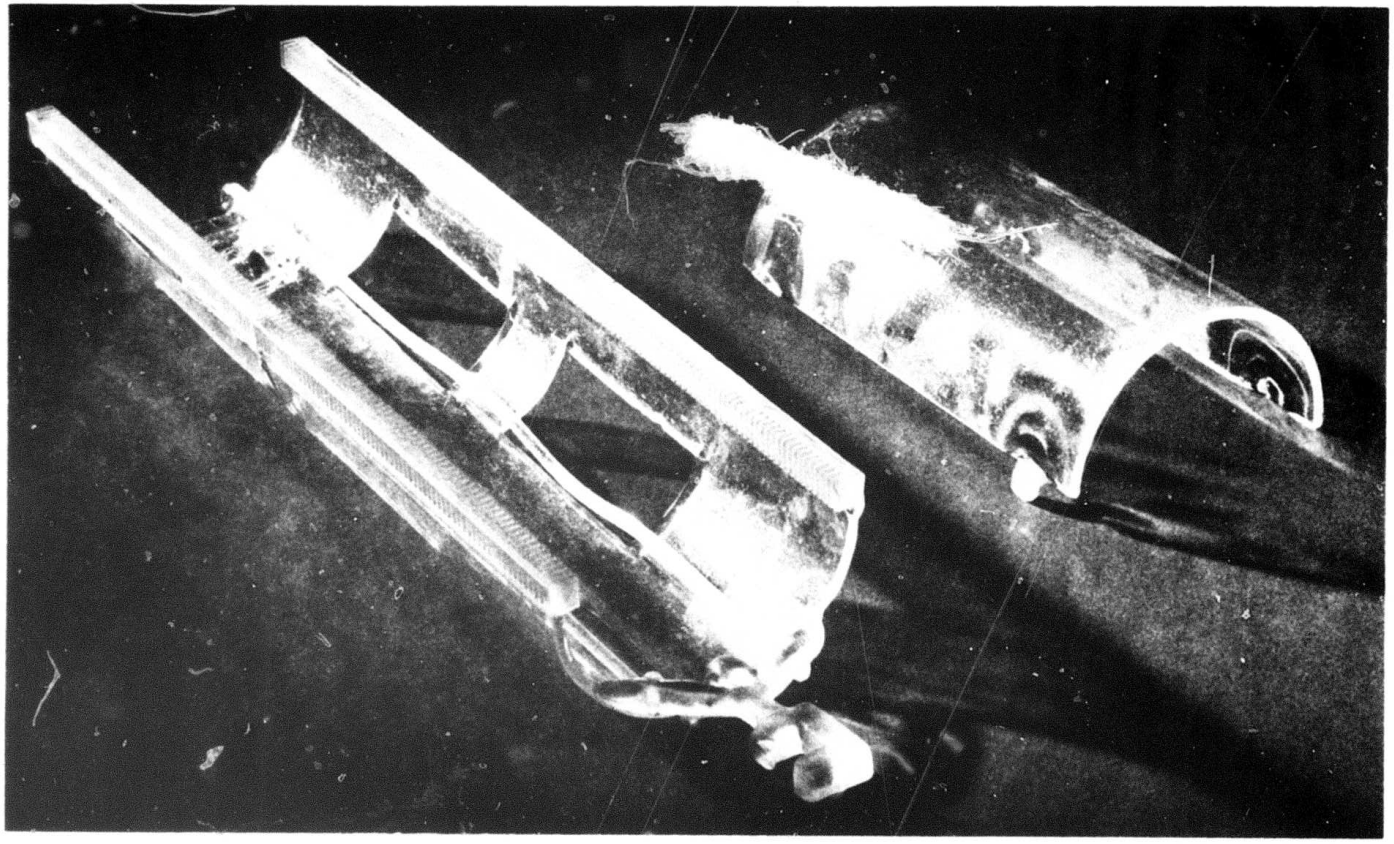
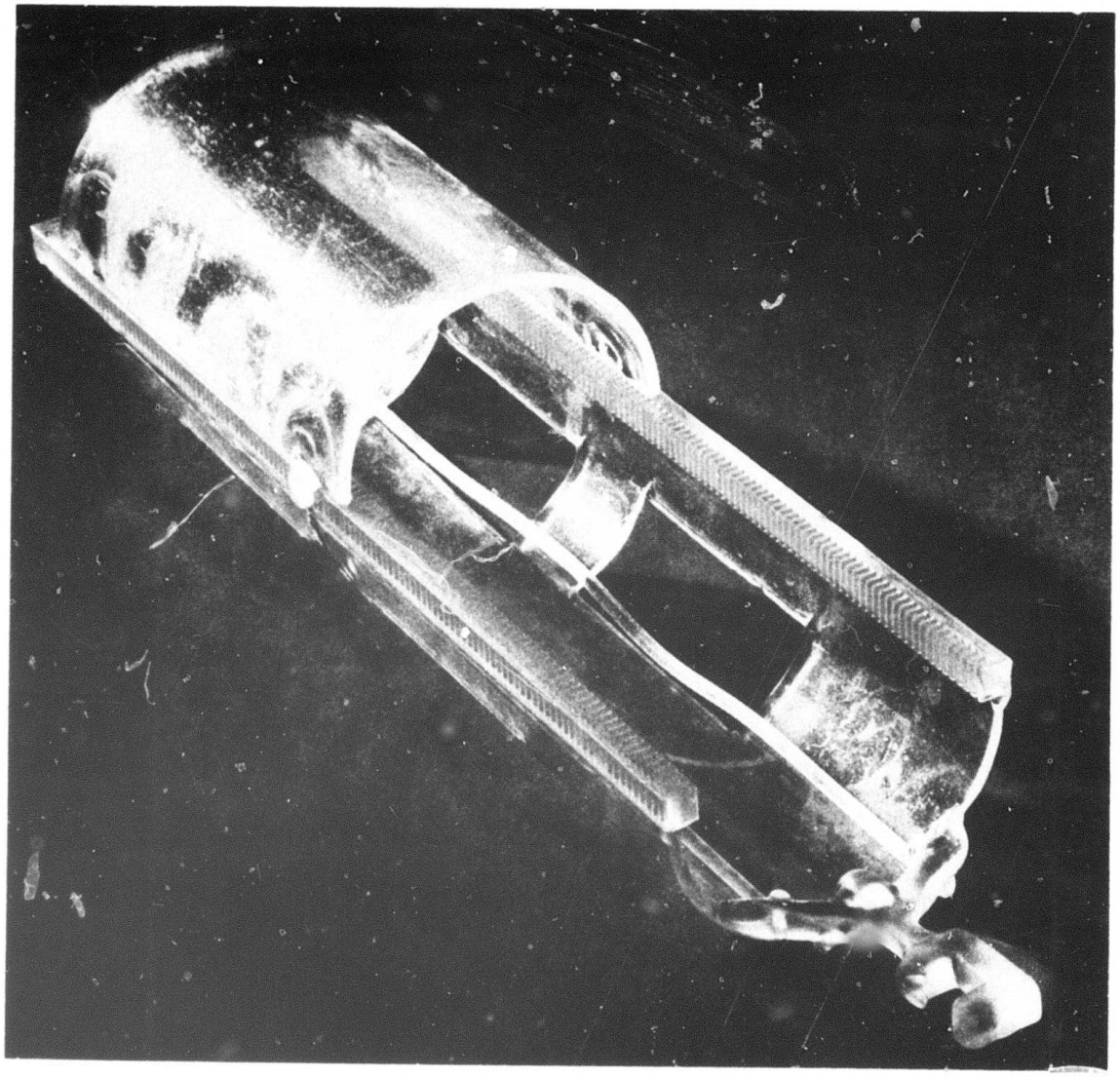


Fig. 1a. Standard quartz boat with cover.



Closed Boat 5

Fig. 1b. Standard quartz boat, one half covered.

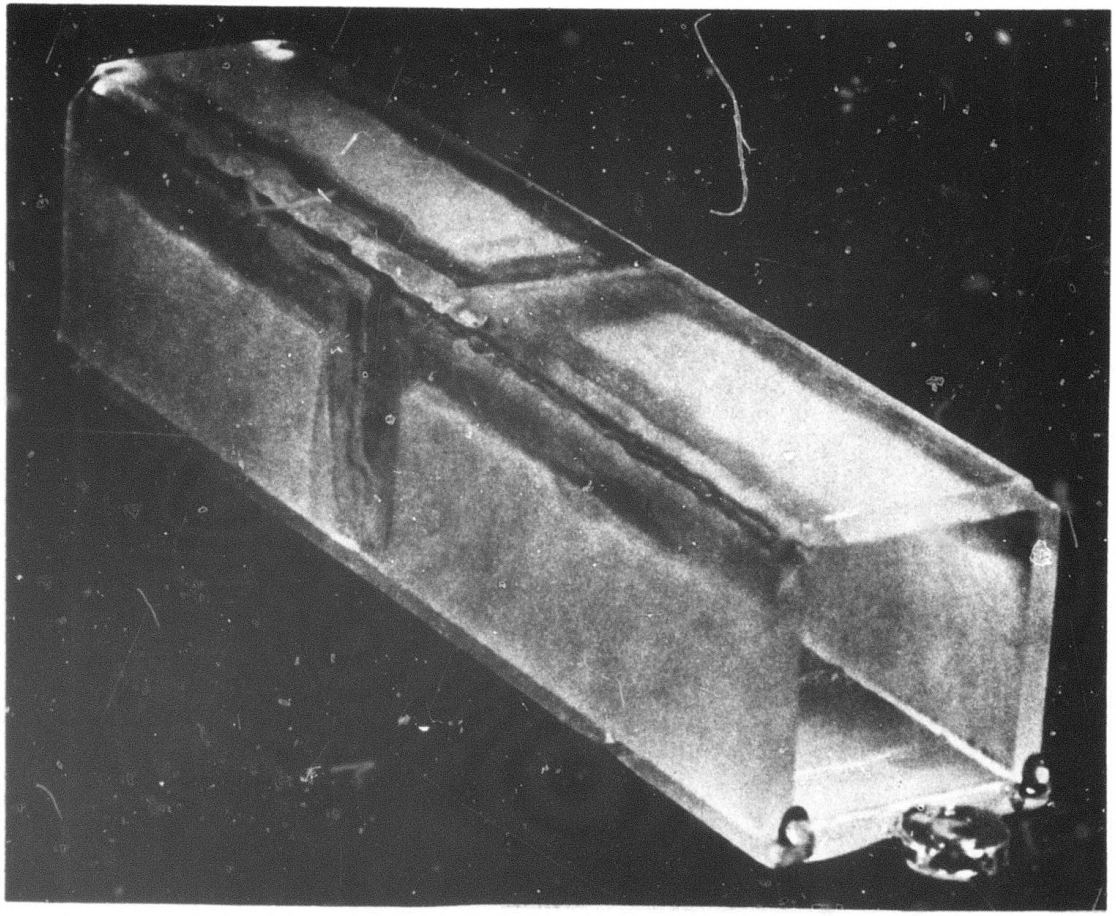
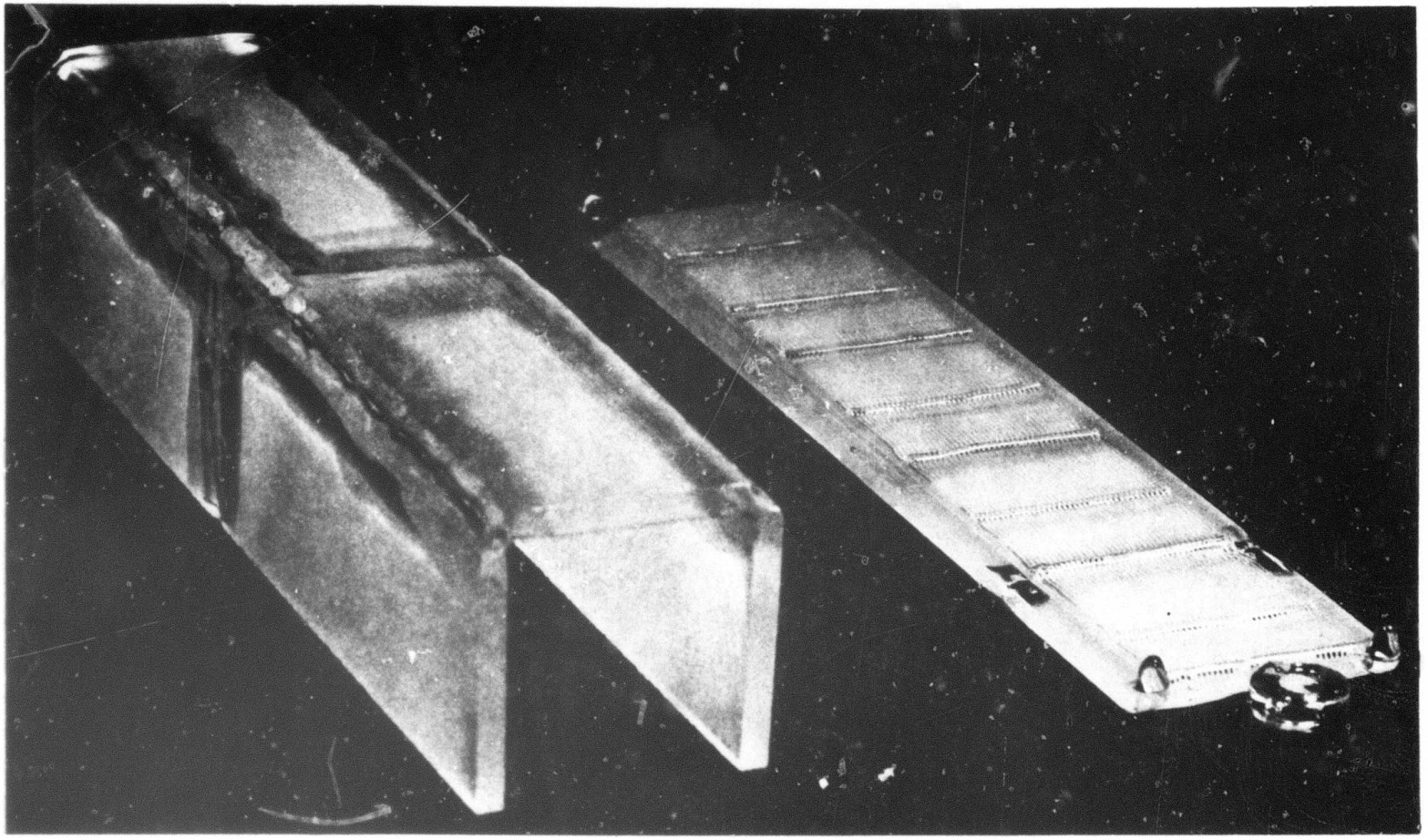


Fig. 1c. Closed quartz boat.



Closed Boat 7

Fig. 1d. Closed quartz boat, uncovered.

temperature furnace and thus cooled to room temperature. The rapid cooling in the initial period is primarily controlled by radiative cooling leading to a radial temperature gradient in the wafers. The wafer periphery cools faster than the wafer center. Thus thermal stresses are set up in each wafer during the cooling period. Frequently, the temperature gradient is large enough to cause plastic deformation of the wafer. Typical examples of dislocation patterns observed in wafers after high temperature processing are shown in the x-ray topographs of Fig. 2. Such peripheral dislocations were previously shown to degrade device yield (1). Direct measurements of temperature gradients responsible for such dislocations are reported in the following. It is also shown that these gradients as well as their impact on crystal perfection can be substantially reduced if a closed boat is used for wafer processing.

3.1 TEMPERATURE GRADIENT MEASUREMENTS

Direct temperature measurements were made on wafer surfaces to obtain quantitative data of actual temperature differences developing between center and periphery of a wafer during heating and cooling cycles. For the measurements two beads of differential thermocouple were positioned such that the

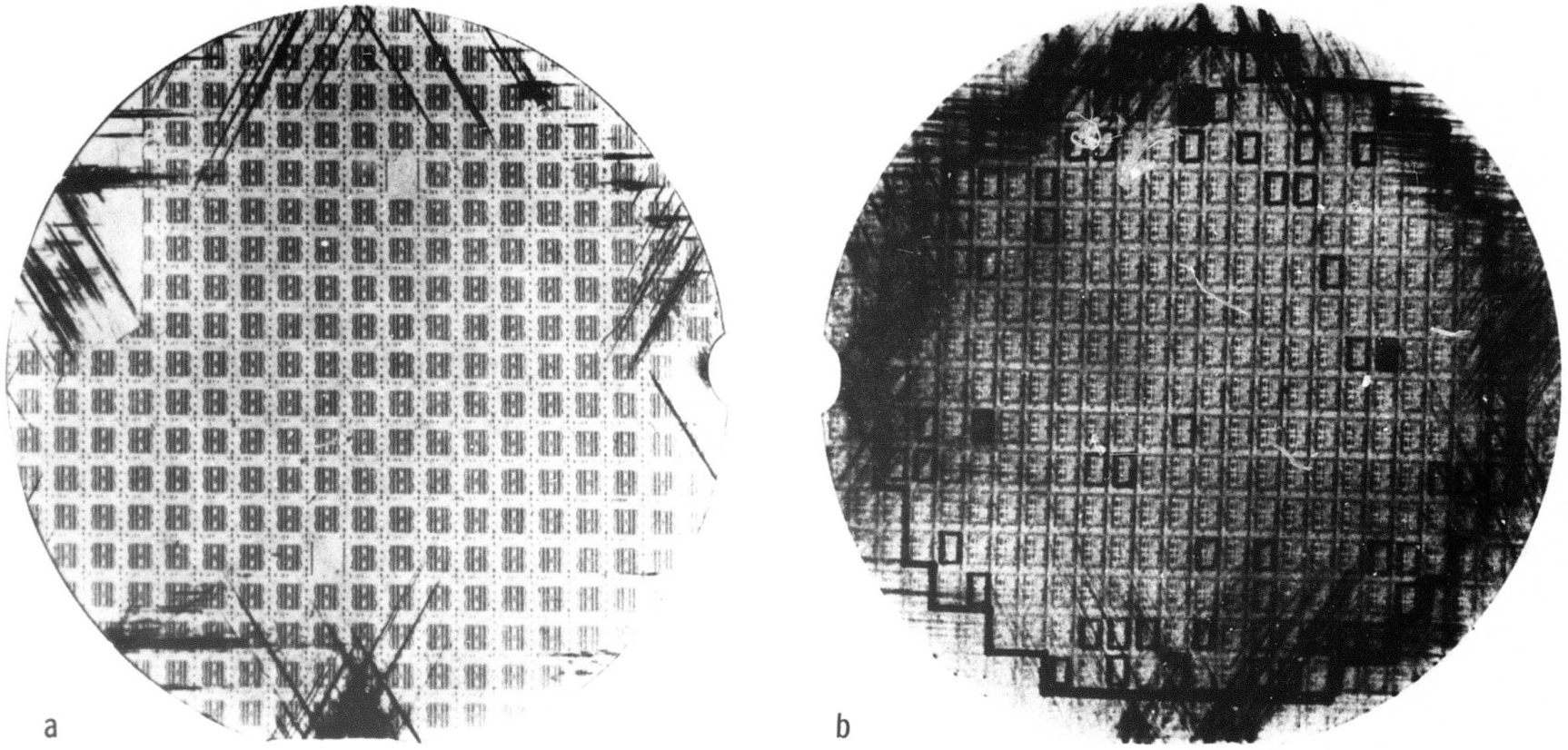


Fig. 2. X-ray topograph of silicon wafers after high temperature processing.

a. After subcollector reoxidation.
b. After emitter diffusion. Yield map superimposed. All peripheral devices outside line failed. Bad devices inside line are also indicated. Note correlation between process induced defects and final device yield.

Closed Boat 9

resulting temperature difference between center and periphery of a wafer could be measured. The output of the differential thermocouple went to a differential voltmeter and from there to a strip chart recorder. A typical output tracing of a heating and cooling cycle of a wafer positioned in the middle of a fully loaded boat is shown in Fig. 3. The horizontal axis in this figure is the time axis in minutes and the vertical axis is the temperature axis in degrees centigrade. The zero or center line represents a zero gradient. The thermal gradient is recorded as the deviation to the plus side of the center line for the outer rim if the wafer is hotter than the center and to the minus side if the converse is true. Two tracings are represented. The upper curve (Fig. 3a) is generated with the beads of the thermocouple not touching the wafer surface but being as close to it as possible. The lower curve (Fig. 3b) is measured with the beads touching the wafer surface. No significant differences are noticed between the two techniques. A slight difference in the output curves is observed for the heating cycle. The cooling cycle output curves are practically the same whether or not the beads are in contact with the wafer surface. In subsequent experiments it was found that such variations in the heating cycle gradients were due to the response time of the total setup consisting of thermocouple beads, wires, boat and

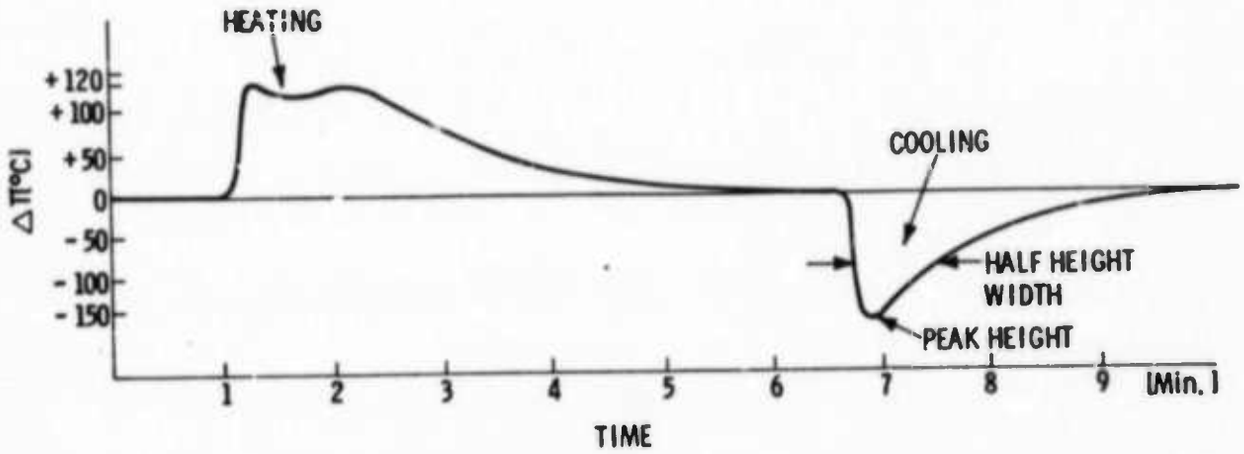


Fig. 3a. Heating and cooling curves -- thermocouple not touching wafer.

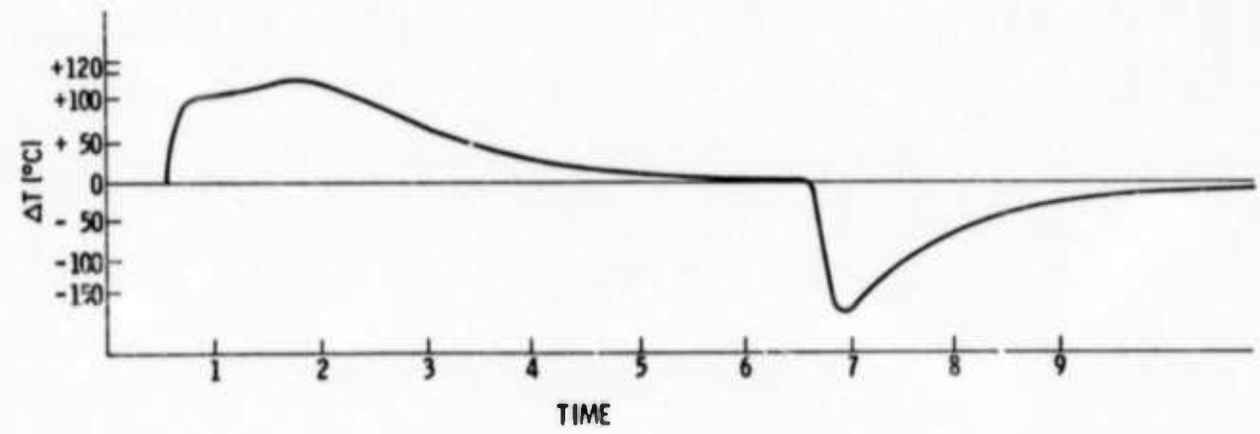


Fig. 3b. Heating and cooling curves -- thermocouple in contact with wafer surface.

Closed Boat 11

wafer load. Since only the cooling cycle is of interest in this experiment, the peak heights of the cooling curves are used as the important parameters. It should be noted that such measurements pose some experimental difficulties. Several configurations of touching, not touching the wafer, drilling holes into the wafer, or using various glues to attach the beads to the surface were tried but were not successful. To prevent the silicon-platinum reaction (at about 500°C) to occur when the thermocouple beads touched the wafer surface the silicon wafers had to be slightly oxidized for such measurements. The best configuration was found to be the one with the beads not touching the wafer surface but being very close to it. The temperature gradient was measured for the two different boat configurations shown in Figs. 1a and 1b. Figure 4 shows the cooling curves for the center wafer of a fully loaded standard boat placed into a furnace at 900°C and withdrawn to air after equilibrium was reached. The thermocouple output of the curve labelled "no top" shows a negative temperature difference of approximately 150°C. This indicates that the centers of the wafers in the stack are hotter than their outer rims. The curve labelled "with top" shows a reduction of the thermal gradient by a factor of 2 1/2 to 3. This curve was generated upon cooling the wafer stack with a top placed on the boat as shown in Fig. 1b. It

was found that the maximum thermal gradient of the center wafer varied with the number of wafers in the stack put into the boat. The experiments showed also that for our boat configurations the thermal gradient approached a maximum temperature for a stack of fifteen wafers.

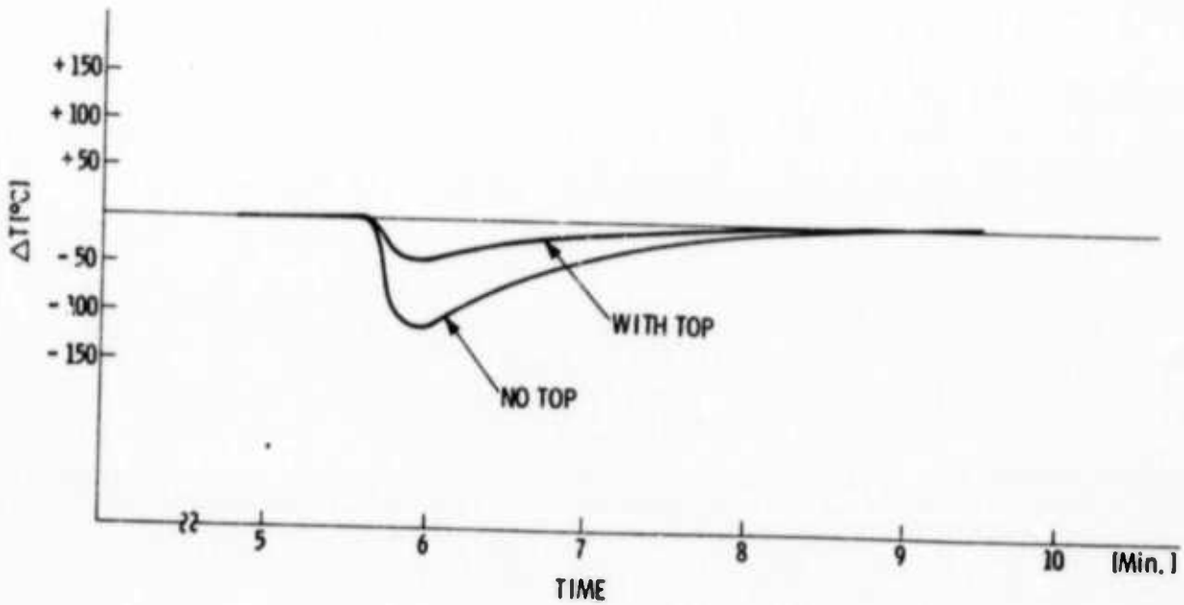


Fig. 4. Cooling curve for standard boat (no top) and covered boat (with top).

3.2 CRYSTAL PERFECTION DURING BATCH PROCESSING

Observation shows that thermal stresses generated in silicon wafers during the cooling period of hot wafers stacked in a row can be large enough to exceed the yield stress of silicon and consequently generate dislocations in the periphery of the wafers (Fig. 2). The magnitude of the stresses can be estimated from the simple expression $\sigma=\alpha E\Delta T$ (Ref. 7), where σ is the thermal stress in the wafer generated during the cooling period as a result of the temperature gradient ΔT. E equals 1.5×10^{12} dyn/cm^2 and is the average Young modulus (8), while α is the coefficient of linear expansion (9) and is approximately 4×10^{-6} K^{-1}. Accordingly a thermal gradient of 150°C is already large enough (0.9×10^{9} dyn/cm^2) to exceed the yield stress (0.45×10^{9} dyn/cm^2) of silicon (10) and to introduce dislocations in agreement with the experimental observation. In practice the situation is aggravated through the occurrence of stress risers as a result of mechanical damage around the periphery of the wafer. Consequently, flats and notches or other peripheral damage (such as small broken out chips as a result of handling) promote the nucleation of dislocations for even smaller temperature gradients. For higher temperature treatments considerably larger ΔT's can arise. As a rule of thumb we find ΔT approximately $0.2T_o$, where T_o

is the temperature of the wafer in the furnace in centigrade before cooling. Exact measurements of the temperature gradient measured for the range 700 to 1000°C are summarized in Fig. 5. The effectiveness of the closed boat (Fig. 1b) in reducing temperature gradients compared to the open boat (Fig. 1a) is evident. A series of x-ray topographs - showing the influence of such gradients on crystal perfection - is given in Fig. 6. The good perfection of the silicon wafers before heat treatment is shown in Fig. 6a. A silicon slice taken from the center of the open boat after a quench from 1100°C to room temperature is shown in Fig. 6b. The wafer periphery is highly dislocated. An x-ray topograph of a similar wafer but heat treated in the closed boat at 1200°C is shown in Fig. 6c. Note that this time the perfection is excellent as compared to the wafer heat treated at only 1100°C in the open boat. However, some dislocations are generated in the vicinity of the wafer flat indicating residual mechanical damage in this area.

A further improvement can be achieved by adjusting the wall thickness of the closed boat. Such a boat is shown in Fig. 1c and excellent results were obtained with it. An example is given in Fig. 7 contrasting two wafers heat treated at 1200°C. The topographs are recorded after a quench from 1200°C to room temperature. Figure 7a shows the

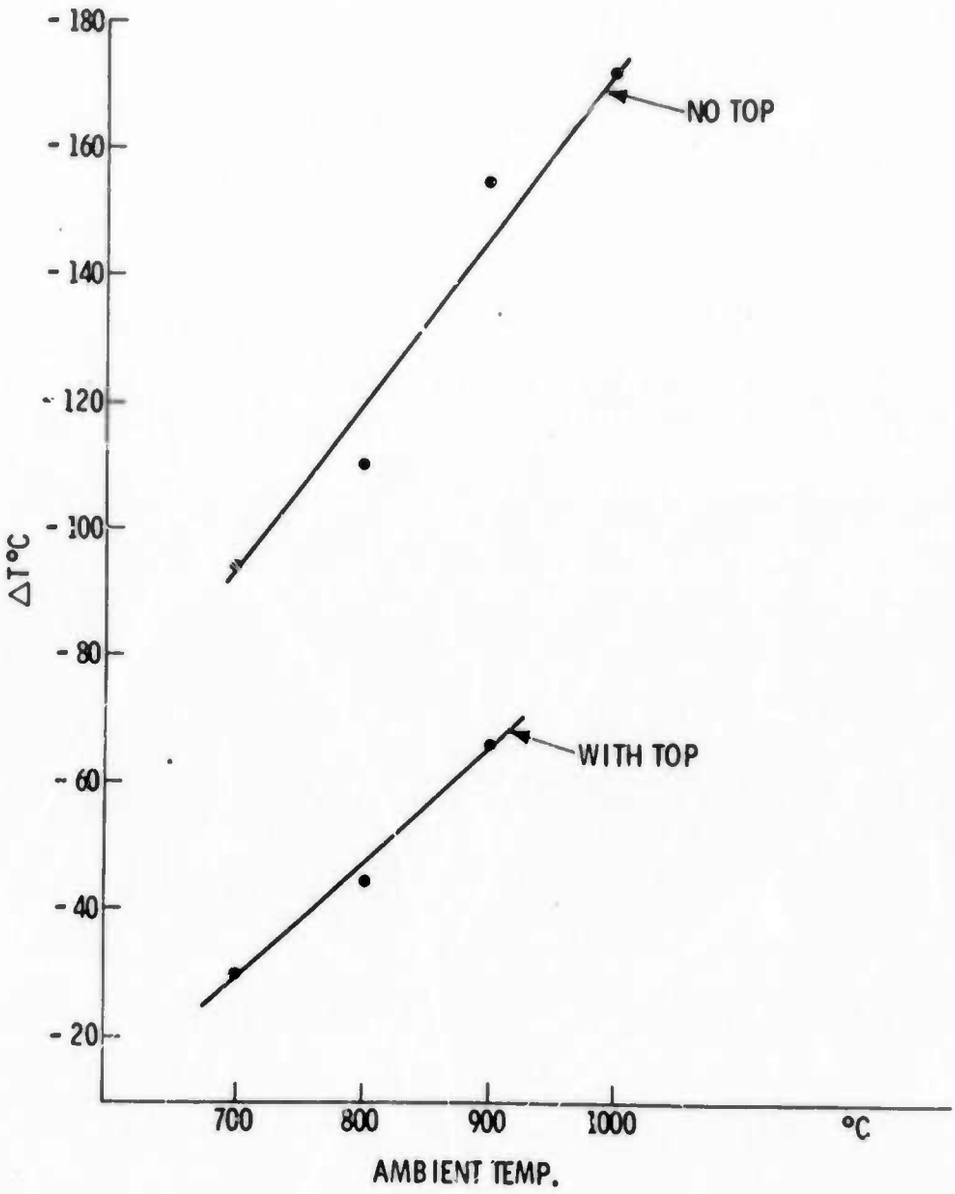


Fig. 5. Temperature gradients during cooling for center wafer measured between 700°C to 1000°C for standard boat (no top) and covered boat (with top).

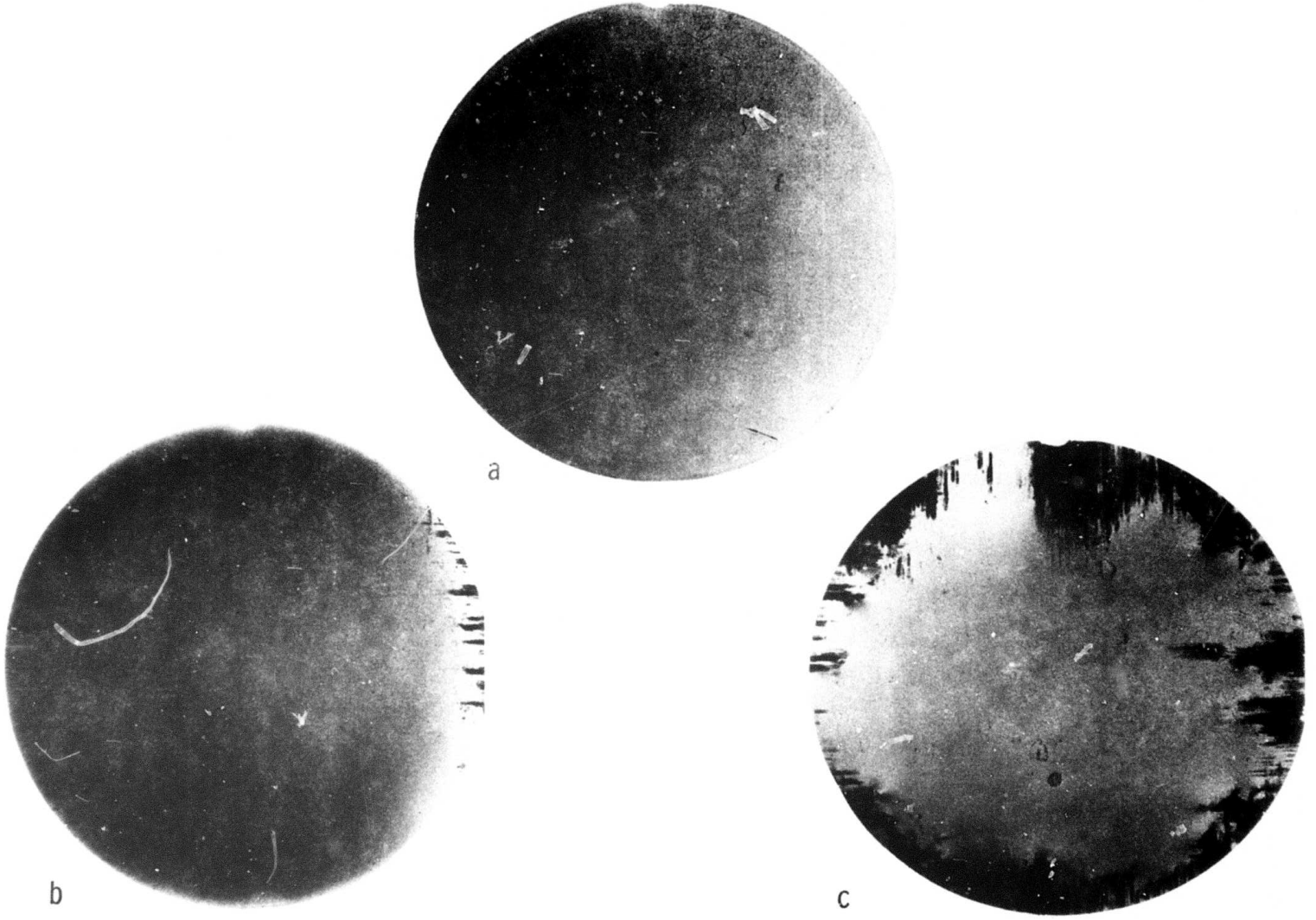


Fig. 6. X-ray topographs of wafers before and after processing.

a. Before heat treatment.
b. Quenched from 1100°C in the open boat.
c. Quenched in the covered boat from 1200°C.

Closed Boat 17

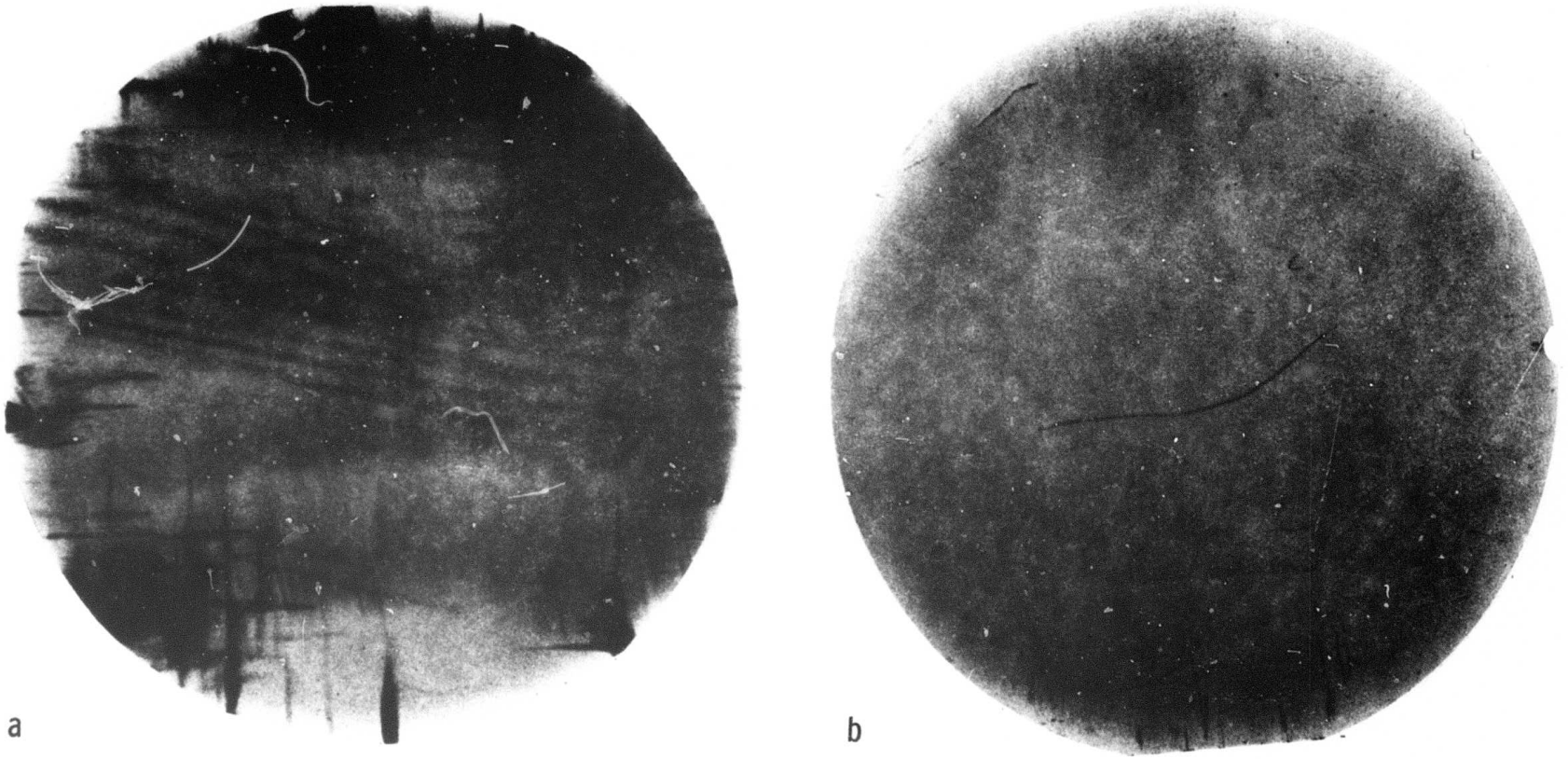


Fig. 7. X-ray topographs of wafers before and after processing at 1200°C.

a. Open boat (composite of several topographs due to large amount of warpage).
b. Closed boat of Fig. 1c.

x-ray topograph of the wafer heat treated in the open boat and Fig. 7b displays the x-ray topograph of a similar wafer heat treated in the closed boat corresponding to Fig. 1c. While the wafer quenched in the open boat shows profuse dislocation generation the wafer processed in the closed boat of Fig. 1c retains its original perfection. Note that even scratches present in the wafer surface generate only very small dislocations in support of the excellent cooling properties of this boat. Noteworthy also is that the topograph, shown in Fig. 7a, is a composite topograph. The wafer contained a considerable amount of warpage in addition to the dislocations as a result of the high temperature processing. Thus it is seen that closed boat processing reduces, or eliminates substantially, warpage due to fast cooling of wafers. This is discussed in more detail in the following section.

3.3 WAFER WARPAGE DURING BATCH PROCESSING

A body of homogeneous material should be free of thermal stress and deformation after passing through a transient temperature cycle as long as its stress-strain relationship stays within the elastic range. As shown in the preceding section thermal stresses set up in wafers during batch processing in a standard boat exceed quite often the yield

stress during cooling. Consequently, plastic deformation occurs in the wafers partially relieving thermal stresses. Since thermal stresses are only partially relieved a reversed stress distribution arises in a wafer after cooling which now - with the wafer at room temperature - cannot be relieved by plastic deformation. The wafer being a thin plate will relieve such strains by buckling or warping. Consequently, a heat-cycled wafer may lose its plane shape which is a requirement for successful photolithographic processing. This is shown very clearly in the composite x-ray topograph of Fig. 7a.

The closed boat is also very successful in reducing such warpage in heat-cycled silicon wafers. This can be seen from the following data. Quantitative warpage data of wafers processed in different boats were obtained by measuring the elevation of a wafer before and after heat treatment above an optical flat. Warpage results are shown for 57.2mm wafers in Fig. 8. In Fig. 8 the warpage is plotted on the vertical axis as a function of wafer position in the boat which is the horizontal axis. The measurements summarized in Fig. 8 compare warpage obtained in the standard boat (Fig. 1a) versus the closed boat obtained by putting a roof on the standard boat (Fig. 1b) and also versus the closed boat corresponding to Fig. 1c. The

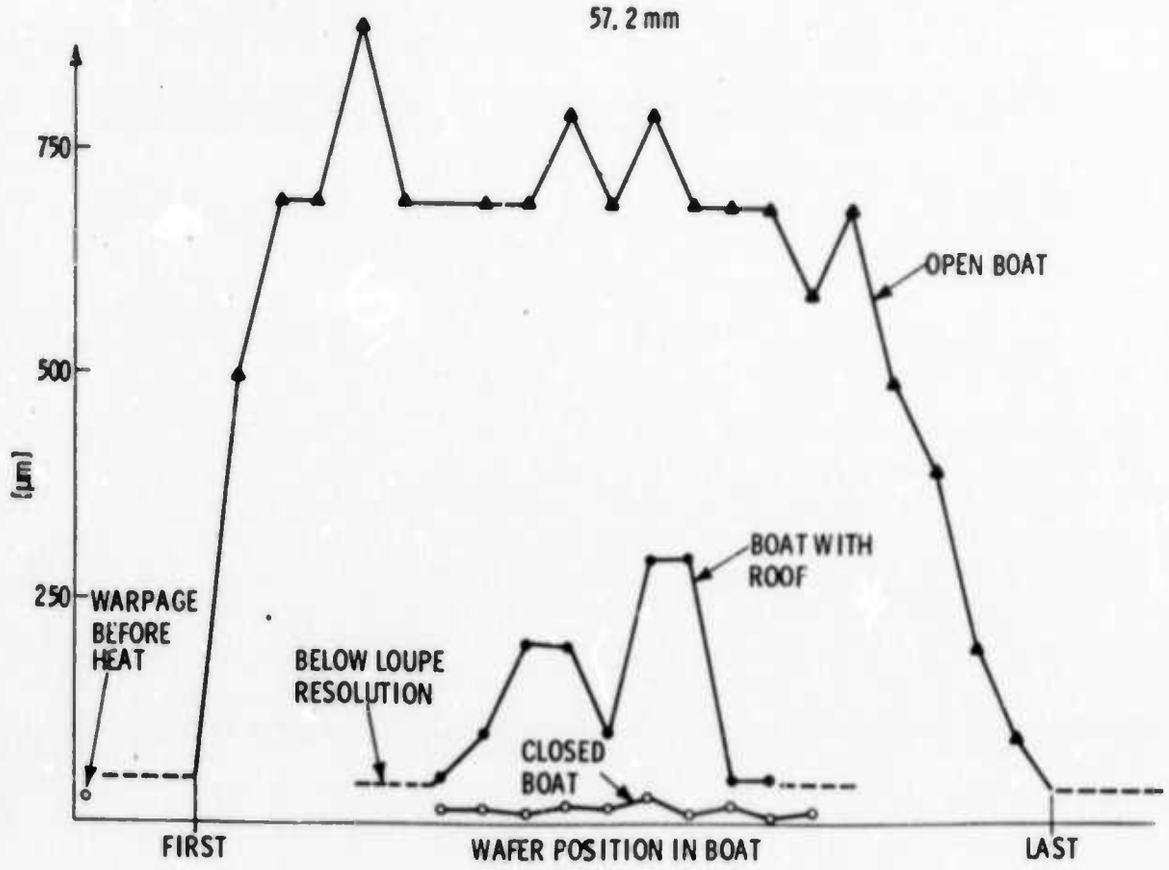


Fig. 8. Warpage measurements for 57.2mm diameter wafers. Note difference in warpage for different boats.

warpage is seen to decrease from a maximum of 750μm obtained in the open boat to approximately 250μm for the boat with roof. This is a decrease by a factor of 3 and identical to the decrease in temperature observed for the closed boat compared to the open boat (Fig. 5). No warpage is observed after processing the wafers in the closed boat shown in Fig. 1c. This indicates that warpage in silicon wafers as a result of heat cycling can be substantially reduced through proper boat selection.

Temperature gradients and their consequences are more severe for larger diameter wafers. This is shown for 75mm diameter wafers and different boats in Fig. 9. Again it is obvious that warpage is substantial in wafers processed in standard boats and that it is reduced through processing in the closed boat. According to Fig. 9 warpage is about four times as large for 75mm wafers as compared to the 57.2mm wafers when processed in the standard boat of Fig. 1a. The closed boat in Fig. 1b reduces the warpage considerably but not sufficiently. Large diameter wafers are more effectively cooled using the thick wall boat shown in Fig. 1c. Similarly good results are obtained if the simple boat shown in Fig. 1b (originally designed for 57.2mm wafers) is pulled rapidly into a furnace end-cap that is wrapped with a stainless steel foil as shown in Fig. 10.

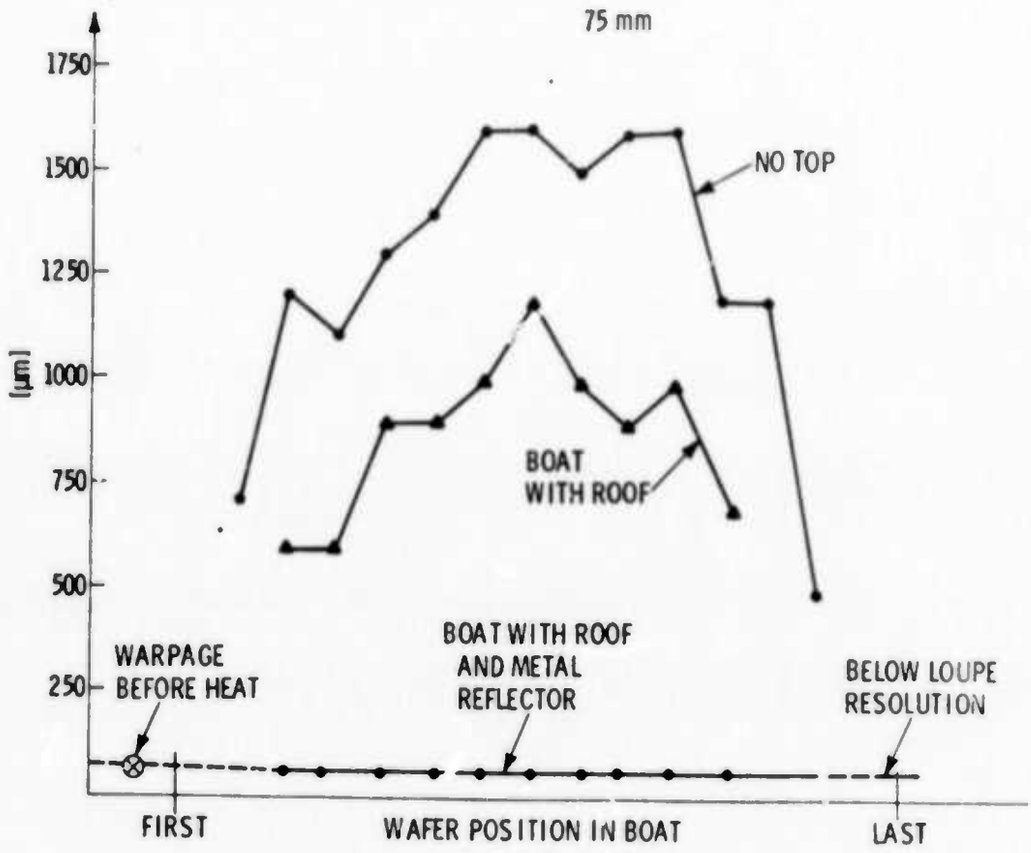


Fig. 9. Warpage measurements for 75mm diameter wafers. Note that warpage is four times as large for 75mm wafers compared to 57.2mm wafers when processed in open boat.

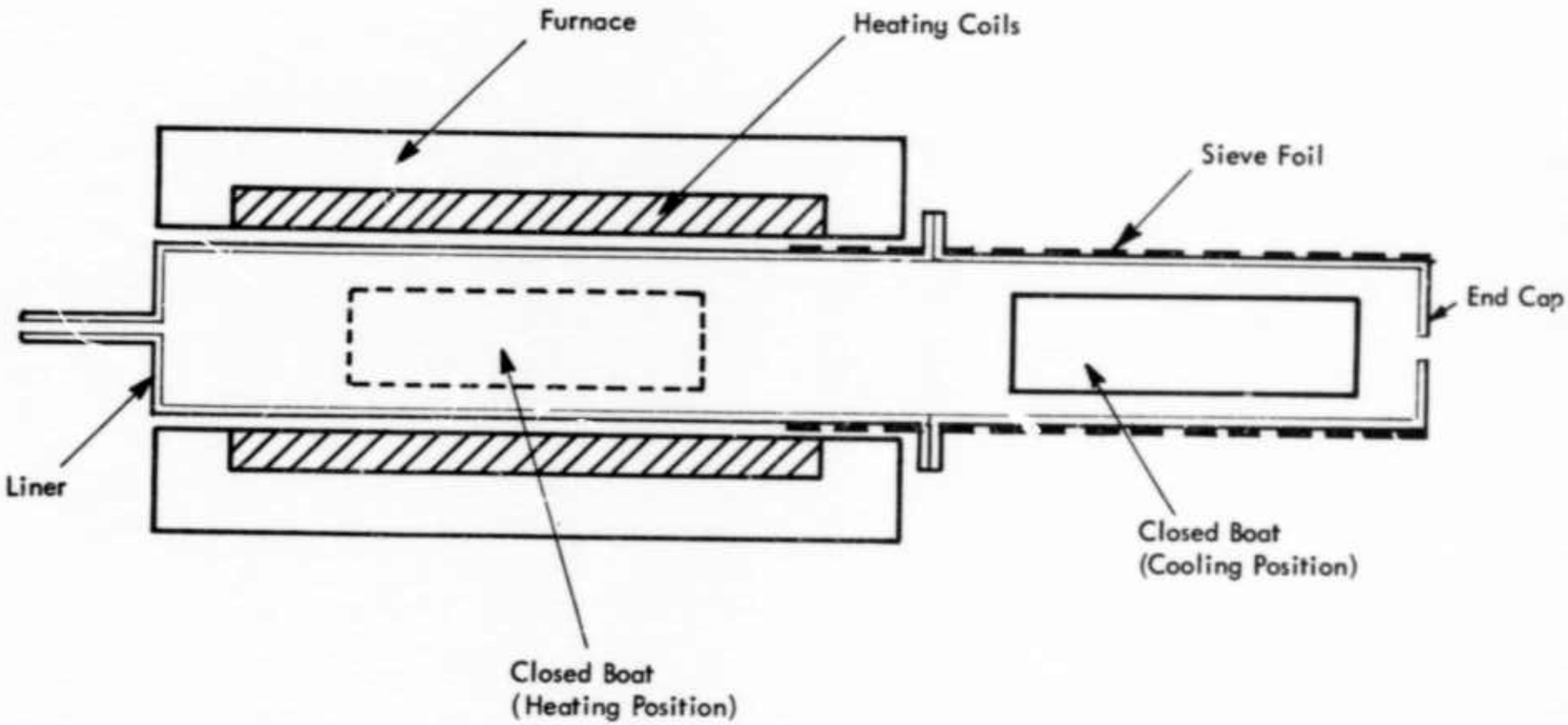


Fig. 10. Schematic of furnace end-cap technique.

Note that all warpage measurements presented in Figs. 8 and 9 relate to heat cycles of 1200°C.

4. BATCH PROCESSING OF SILICON SLABS

Presently square or rectangular shaped silicon wafers are not used routinely in semiconductor manufacturing except for solar cell fabrication. The processing of large rectangular wafers - slabs - can pose some severe temperature problems if the open boat is used but such problems disappear in the closed boat. This is shown for slabs of 115mm and 57mm and of 0.375mm thickness. The slabs were cut with the wafer surface of <110> orientation. The wafer surfaces were prepared using the same chemical-mechanical polishing technique as for 57.2 and 75mm diameter slices. The slabs were loaded into boats with their long axis parallel to the boat axis. For open boat processing 15 wafers were standing side by side spaced 3mm apart. The closed boat was of the type shown in Fig. 1c. The bottom part of this boat contained also 15 slots, each at a distance of 3mm for slab loading. X-ray topographs were made of the slabs before and after heat treatment. Typical examples of wafer perfection before and after processing are shown in Fig. 11. The advantage of closed boat processing compared to open boat processing is also supported through this experiment.

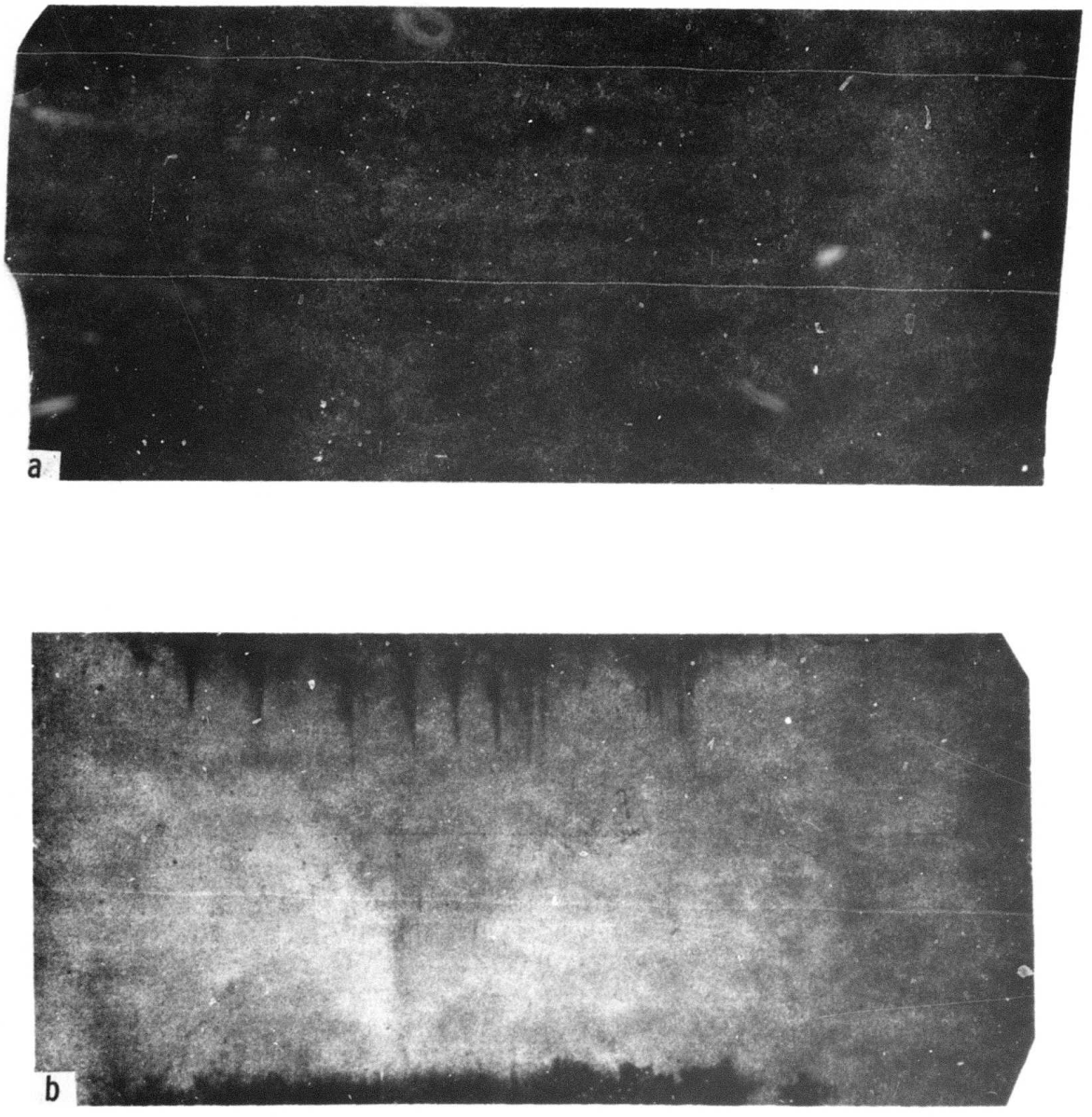
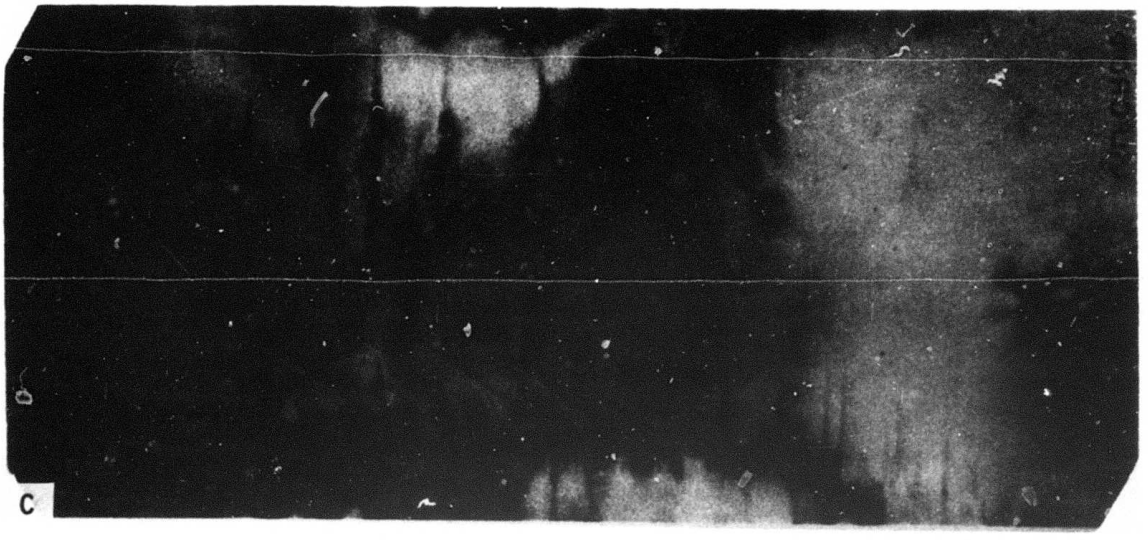


Fig. 11. X-ray topographs (recorded on x-ray film due to large size) of large slabs after heat treatment. Note perfection resulting through use of closed boat.

a. Before processing.
b. After quench from 1100°C, open boat.



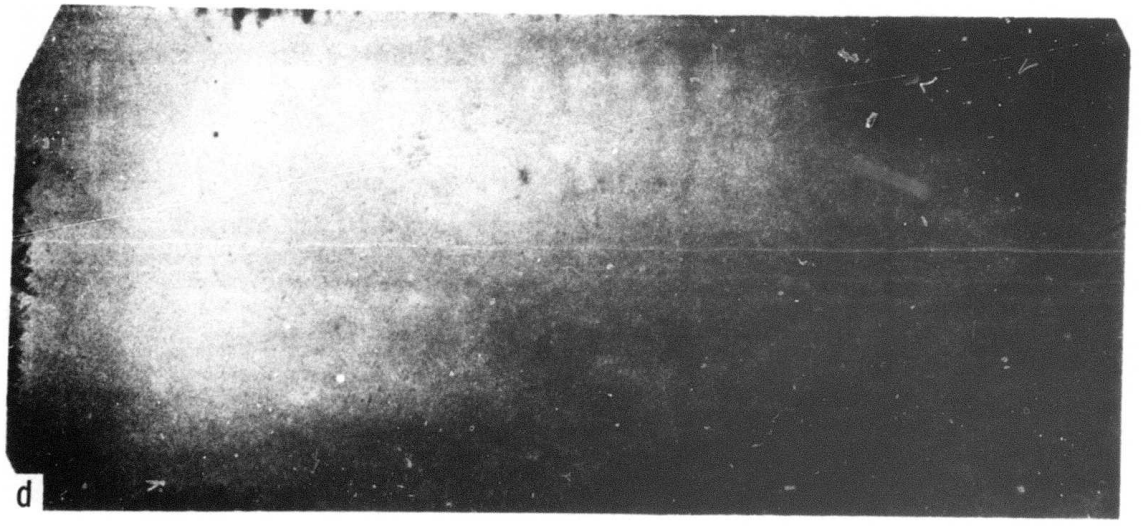


Fig. 11. X-ray topographs of large slabs after heat treatment.

c. After quench from 1200°C, open boat.
d. After quench from 1200°C, closed boat.

Closed Boat 27

5. DISCUSSION

Measurements reported in this paper demonstrate that a row of regular spaced wafers (batch) stacked in a boat is susceptible to drastic temperature changes whenever the boat is quickly pulled out of a furnace at high temperature and cooled to room temperature. Such changes can be reduced if a closed boat is used. The closed boat is very simple in design and almost trivial in its implementation. In its simplest form the closed boat uses a quartz tube cut in half and put on the standard boat as a roof. A further improvement of this simple approach is the use of the furnace end-cap wrapped with a metallic foil (Fig. 10). The reflectivity of such a foil can be adjusted by perforating the foil. The combination of simple closed boat and wrapped furnace end-cap technique can be used to achieve maximum cooling rates for any wafer size and load without introducing crystal defects or warpage into the wafers.

To overcome yield losses due to high temperature induced slip, the semiconductor industry has generally accepted slow cooling of wafers. Slow cooling of wafers requires fairly sophisticated and automated equipment. The slow cooling technique is also at a disadvantage compared to the closed

boat when sharp furnace gradients terminate the flat zone of the furnace. Measurements on several empty furnaces indicated to us that quite often gradients larger than 200°C accompany the termination of a flat zone in a furnace. In a slow cooling process - particularly with the standard boat - it is no surprise that such gradients are felt by the wafers when the boat is programmed to move slowly at a predetermined rate out of the furnace. Contrary, the closed boat is pulled rapidly within a few seconds out of the furnace and consequently it is not influenced by such intrinsic gradients.

It is also our experience that primarily the cooling cycle is detrimental to wafer quality. This may seem surprising because the wafers experience a similar gradient whenever they are pushed into the furnace (Fig. 3). However, this effect can be understood by realizing that during heating the wafers experience the gradient at a lower temperature than in the cooling cycle. In addition, it is necessary to realize that similar temperature gradients cause different stresses depending on the direction (sign) of the gradient (Fig. 3). Stresses in a thin plate due to a radially dependent temperature gradient almost vanish for hot edges and cool center because under such conditions a plate can freely expand (11).

Mechanical damage around the wafer periphery such as chips, broken etches, flats and notches, and surface scratches occur always as a result of wafer handling. Such damage causes a localized stress in the wafer. The additional stress necessary to generate dislocations in this area is now smaller by the stored amount. The result is normally a preferential generation of dislocations in mechanically damaged areas. Such dislocation sources are also effectively controlled through the closed boat technique. An excellent example of such control is shown in Fig. 7b.

One of the most striking observations about peripheral dislocation generation as a result of open boat processing may raise the question why the bottom of the wafer - which is protected by the boat - shows as many defects after cooling as the upper part of the wafer - which is not protected (Fig. 2). The answer is obviously that the peripheral stress is the same around the circumference of the wafer and, consequently, the defect pattern in the wafer reflects the symmetry (orientation) of the wafer. This is clearly seen in the topographs of Fig. 2. Preferential dislocation generation is initiated in damaged areas.

According to our measurements wafer warpage is another

disadvantage of the open boat batch processing and likewise a function of temperature gradients generated during the cooling cycle just like dislocation generation. An additional parameter that can affect wafer warpage is the surface quality of the wafer. Chemical-mechanical polished wafers showed less tendency to warp than mechanical polished wafers. Differently polished front and back surfaces on one wafer were also found to influence wafer flatness adversely in the open boat processing. Again we found that such parameters are favorably influenced through the closed boat technique.

6. SUMMARY

Temperature gradient measurements and x-ray topography are used to characterize silicon wafers processed in a row (batch) through high temperature cycles. Experiments are performed with the furnace at 900°C, 1100°C and 1200°C. For processing, the wafers are loaded in standard and modified quartz boats. Temperature gradients that occur in wafers during the heating cycle while the boat is pushed rapidly into the hot zone of the furnace and during the cooling cycle while it is rapidly pulled out of the furnace are measured and related to crystal perfection in the wafers. It is shown that a closed boat can reduce such gradients and thus substantially improve cyrstal perfection in heat-cycled wafers.

REFERENCES

1. G. H. Schwuttke, Microelectronics and Reliability, 9, 397. (1970).

2. J. M. Fairfield and G. H. Schwuttke, J. Electrochem. Society, 113, 1229, (1966).

3. S. M. Hu, J. Appl. Phys. 40, 4413, (1969).

4. G. H. Schwuttke, Proc. 1969 IEEE Annual Symp. Reliability Physics, page 274, (1969).

5. E. W. Hearn, G. H. Schwuttke, E. H. teKaat, U. S. Patent No. 3.737.282, June 5 (1973).

6. G. H. Schwuttke, J. Appl. Phys. 36, 2712 (1965).

7. K. Morizane and P. S. Gleim, J. Appl. Phys. 40, (1969).

8. E. M. Sparrow, Ph.D. Thesis, Harvard University, Cambridge, Mass. (1956).

9. L. Maissel, J. Appl. Phys. 31, 211, (1960).

10. J. R. Patel, A. R. Chaudhuri, J. Appl. Phys. 34, 2788, (1963).

11. C. T. Wang, Applied Elasticity, McGraw Hill Pub. Co., p. 54, 70 (1953).

Chapter 2

A NEW FAST TECHNIQUE FOR LARGE-SCALE MEASUREMENTS OF GENERATION LIFETIME IN SEMICONDUCTORS

by

W. R. Fahrner and C. P. Schneider

INTRODUCTION

The determination of lifetime in silicon (Si) is of great technological importance. Since the basic paper by Zerbst,[1] important contributions to the literature have been made by others.[2-4] Large-scale measurements of minority carrier lifetime have become desirable for many applications. The usual technique, described by Zerbst, is too cumbersome for large-scale measurements. Faster techniques[5] very often do not give the same results as a Zerbst plot. This paper describes a technique capable of rapid measurements of generation lifetime with good precision. Three slightly modified measurement setups using this technique allow coverage of the wide spectrum of lifetimes in silicon which are of practical interest. The measurements are carried out on metal oxide

semiconductor (MOS) capacitors. (We used thermally grown oxides 1000 to 5000 angstrom (Å) thick. The ⟨100⟩ oriented substrates (mostly p-type) had resistivities of 1 to 20 ohm-centimeter (Ω-cm). Aluminum dots were evaporated onto the oxide to form MOS capacitors.) A voltage, V_a, is applied to a metal dot, and a steady-state inversion regime is established. Then a voltage step or pulse, ΔV_a, is added. This creates a depletion layer underneath the dot, thus reducing the value of the total measured capacitance and increasing the (absolute) value of the voltage drop across the interface. The time difference, t, between switching (t=0) and reaching a preselected percentage of the equilibrium value is printed or displayed on a counter. This value is fitted to the thermal generation model by computer.

THEORETICAL

In the theoretical analysis, we write the applied voltage as the sum of the voltage drop, V_{ox}, across the oxide and the surface potential, ϕ_s:

$$V_a + \Delta V_a = V_{ox} + \phi_s. \qquad [1]$$

Differentiation with respect to time and the use of equations

$$\phi_s = q \cdot N_A \cdot \epsilon_{si}/(2\, C_D^2), \qquad [2]$$

$$V_{ox} = (Q_D + Q_I)/C_{ox}, \qquad [3]$$

and

$$dQ_D/dt = C_D \cdot d\phi_s/dt \qquad [4]$$

yields

$$(q \cdot \epsilon_{si} \cdot N_A/2)\, d(1/C_D^2)/dt$$
$$+(C_D \cdot q \cdot N_A\, \epsilon_{si}/2 \cdot C_{ox}) \cdot d(1/C_D^2)/dt$$
$$+(1/C_{ox})\, d\,Q_I/dt = 0. \qquad [5]$$

The integration of Eq. [5] requires knowledge of the generation current, $I_{Gen} = dQ_I/dt$. Most commonly, the thermal generation model

$$dQ_I/dt = q\, n_i (x_D - x_{Df})/2\tau = I_{th} \qquad [6]$$

is adopted. This model is not always valid. Discrepancies are found, expecially for long-lifetime silicon at the end of the C-t return curve, where a faster generation mechanism might exceed the decreasing thermal generation. The origins for these deviations may be channel injection, enhanced generation due to inhomogeneities, or defects in both the oxide and the silicon.

In the case of channel injection or injection due to inhomogeneities, a simple model for dQ_I/dt can be assumed, namely,

$$dQ_I/dt = I_{Gen} = I_{th} + \alpha I_i,$$

$$\text{where } I_i = \begin{cases} I_o \text{ for } C < C_f \\ 0 \text{ otherwise} \end{cases}$$

Note that Eq. [6] does not contain any surface contribution $I_s = n_i \cdot q \cdot s_o$.

In a comparison of I_s and I_{th}, s_0 values in the order of 10 to 700 centimeters per second (cm/sec) were reported.[6] These data are obtained from techniques based on switching the MOS structure from accumulation to deep depletion. By switching from inversion to deep depletion, as done in this technique, one obtains s_0 values in the order of 10^{-2} cm/sec. The reason for the reduction is the screening of the surface by an inversion layer. The lifetimes are found to be in the range of 1 nanosecond (nsec) to 1 millisecond (msec). The space-charge width, x_D, depends on the doping concentration, N_A, and the applied voltage. Assuming typical values $x_D = 10^{-4}$ cm $\gg x_{Df}$, $s_0 = 10$ cm/sec, and $\tau = 50$ microseconds (μsec), one obtains $I_s/I_{th} = 10$, whereas, with the "inversion screened" s_0 value of 10^{-2} cm/sec, I_s is small compared with I_{th}, even for very long lifetimes and small space-charge widths.

In this paper, we adopt the thermal generation model. The reason is shown below, where Eq. [5] is solved for different models for the generation current dQ_I/dt. Combining Eqs. [5] and [6], we obtain

$$\frac{C_{ox} + C_D}{C_{Df} - C_D} \cdot \frac{1}{C_D^2} dC_D = \frac{n_i}{2\tau N_A} \frac{1}{C_{Df}} dt. \qquad [7]$$

This equation can easily be solved by normalization (replacing C_D by $C_D/C_{Df} = C_{DR}$, C_{ox} by C_{ox}/C_{Df}, and C_{Df} by 1) and by changing twice the variable $C_{DR} = 1/x_R = 1/\sqrt{\phi_R}$, where $x_R = x_D/x_{Df}$ and $\phi_R = \phi_s/\phi_{sf}$. The final result is

$$\sqrt{\phi_R} + (C_{Df}/C_{ox} + 1)\ln(\sqrt{\phi_R} - 1) = -t'_R, \qquad [8]$$

where

$$t'_R = \frac{n_i}{2 N_A \tau} \cdot \frac{C_{Df}}{C_{ox}} (t - t_o).$$

This is the basic equation for the measurement of the short lifetimes (low-frequency measurement).

When the MOS capacitance is measured with a high-frequency signal, the total capacitance is $C = C_{ox} C_D/(C_{ox} + C_D)$. In this case, Eq. [7] can be rewritten

$$\frac{1}{C_R^2 (1 - C_R)} \frac{dC_R}{dt} = \frac{C_f}{C_{ox}} \frac{n_i}{2\tau N_A}, \qquad [9]$$

and its solution is

$$\left[\ln \frac{C_R}{1 - C_R}\right] - \frac{1}{C_R} = t''_R. \qquad [10]$$

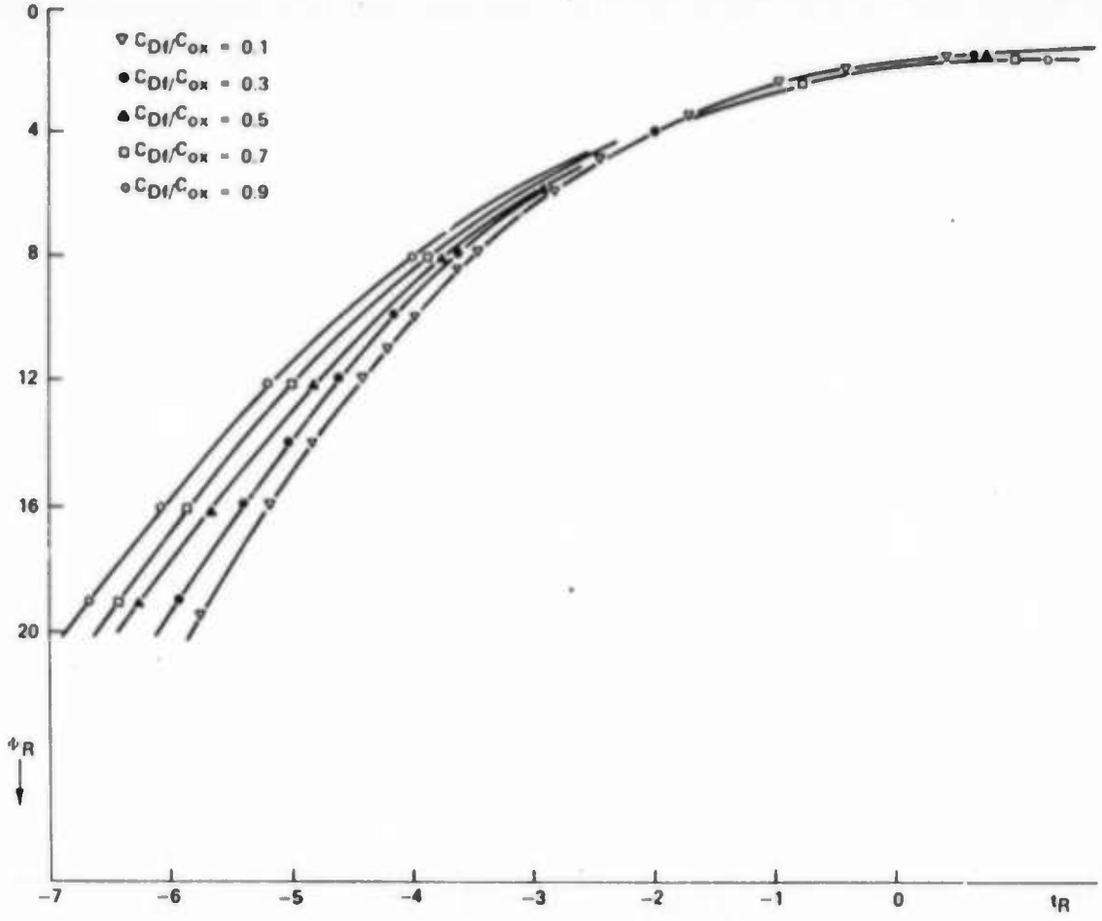


Fig. 1. Relaxation of the surface potential ϕ_R for different values of C_{Df}/C_{ox}. ▽, $C_{Df}/C_{ox} = 0.1$. ●, $C_{Df}/C_{ox} = 0.3$. ▲, $C_{Df}/C_{ox} = 0.5$. □, $C_{Df}/C_{ox} = 0.7$. ○, $C_{Df}/C_{ox} = 0.9$.

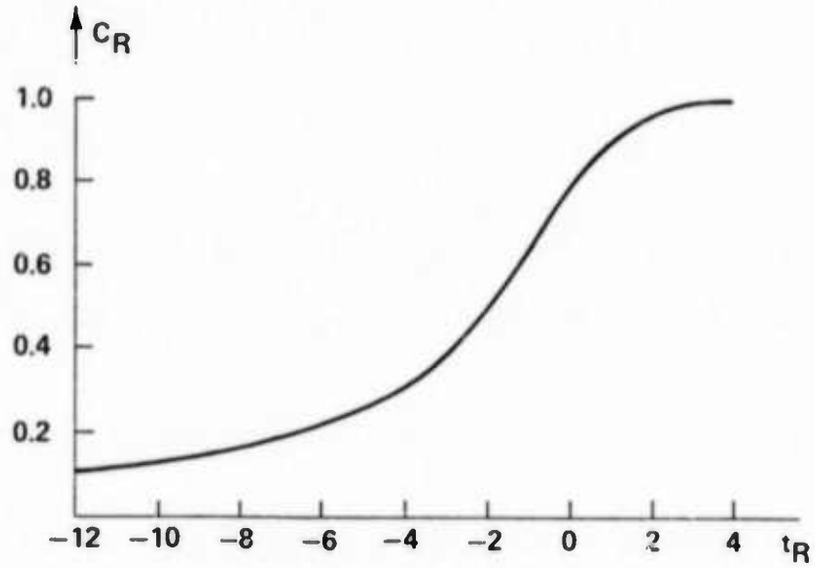


Fig. 2. Relaxation of an MOS capacitance already in inversion to which a voltage step is applied.

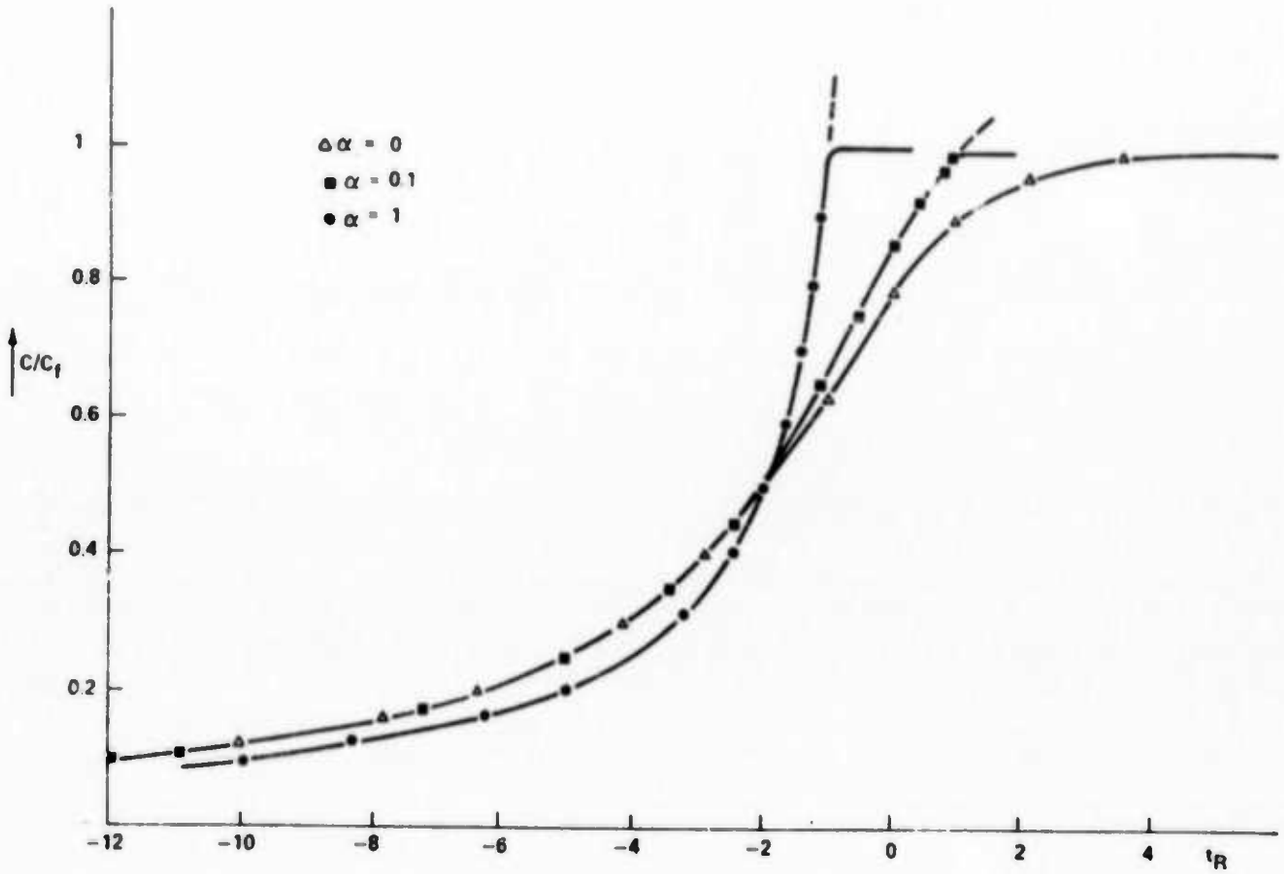


Fig. 3. Relaxation of the capacitance for $I_{Gen} = I_{th} + \alpha I_o$. $\triangle \alpha = 0$. ■$\alpha = 0.1$. ●$\alpha = 1$.

where C_R is the normalized capacitance, $C_R = C/C_f$. Equation [10] is essentially the same as Heiman's result.[5] Heiman, however, did not use this result, but its differential form (Eq. [9]). Furthermore, he switched from accumulation to inversion. For silicon with lifetimes $\tau \geq 10$ μsec, a Heiman plot gives satisfactory results only during a very short range of the measuring time, as shown below. Equation [10] is used here to measure longer lifetimes (high-frequency measurement). Both Eqs. [8] and [10] are plotted in Figs. 1 and 2. In Fig. 1, we reverse the ϕ_R axis because we do the same in the measurement (cf Eq. [12a], below). Equation [5] has been integrated for several other modes of relaxation. Of these, we present one example (Fig. 3), where we add a constant current to the thermal current. The parameter, α, is a measure for the contribution of these additional currents; $\alpha = 0$ means a pure thermal generation. It can be seen that for $C_R \leq 0.9$, if the t_R axis is expanded or compressed, the curves will be similar. On the other hand, a variation of τ has the same effect. Thus, an additional generation current is equivalent to a reduction of the lifetime in the thermal generation model. This observation is valid for any reasonable generation current that decreases towards zero when the capacitance approaches C_f.

We use Eq. [10] (high-frequency measurement) to demonstrate this technique. When we measure two capacitance values $C_R(t)$ and $C_R(t_o)$ at the corresponding times t and t_o, we can calculate τ, because the other parameters involved are easily available. C_{ox} and C_f can be obtained by standard C-$\dot{V}$ measurements.

From these values, N_A can be calculated. n_i has a room temperature value of 1.4×10^{10} cm^{-3}. For t_o we choose the onset of the step or pulse ($t_o = 0$). Rather than measure the capacitance after a fixed time t, we select a fixed capacitance value, e.g., $C_R = 0.8$, and measure the time, t, between switching and reaching the selected capacitance level. Figure 4 shows a schematic C-t plot and the measured time interval.

An analog technique is used in the low-frequency case. Instead of C_R, a specific change of ϕ_R (and its time duration) is measured. Only the counting of the time is slightly different; for example, we count now between $t = 0$ and the time when $\phi_R = 0.2 \times (\phi_R(0) - 1)$ (Fig. 5). This alteration is caused by the experimental conditions (see below).

EXPERIMENTAL AND RESULTS

As shown in Figs. 6a and 6b, the first, high-frequency, measurement setup essentially consists of a 1-MHz capacitance meter, two power supplies, two memories, a voltage comparator, a plotter, and a printer. Care must be taken that one does not switch the voltage through zero, because the majority carrier response is much faster than any switching time. This setup can be used for relaxation times $T = 2\tau N_A/n_i \geq 100$ msec. This lower limit is caused by the transient time of the bridge.

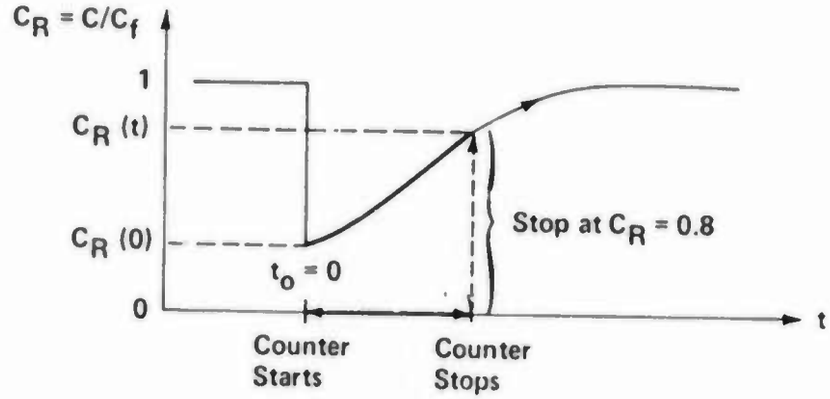


Fig. 4. Representative C-t plot. $t_o = 0$; $C_R = C/C_f$.

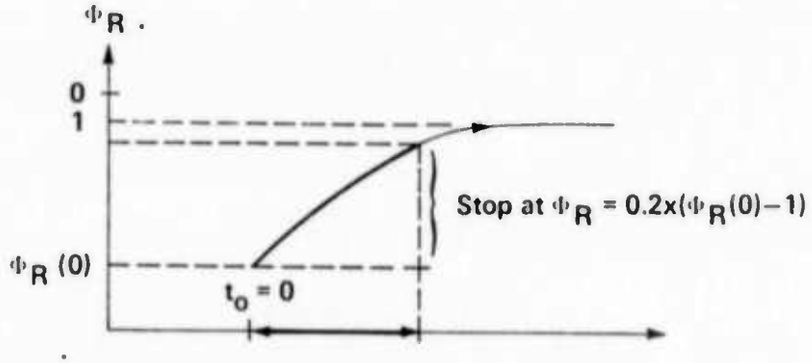


Fig. 5. Representative ϕ-t plot. $t_o = 0$.

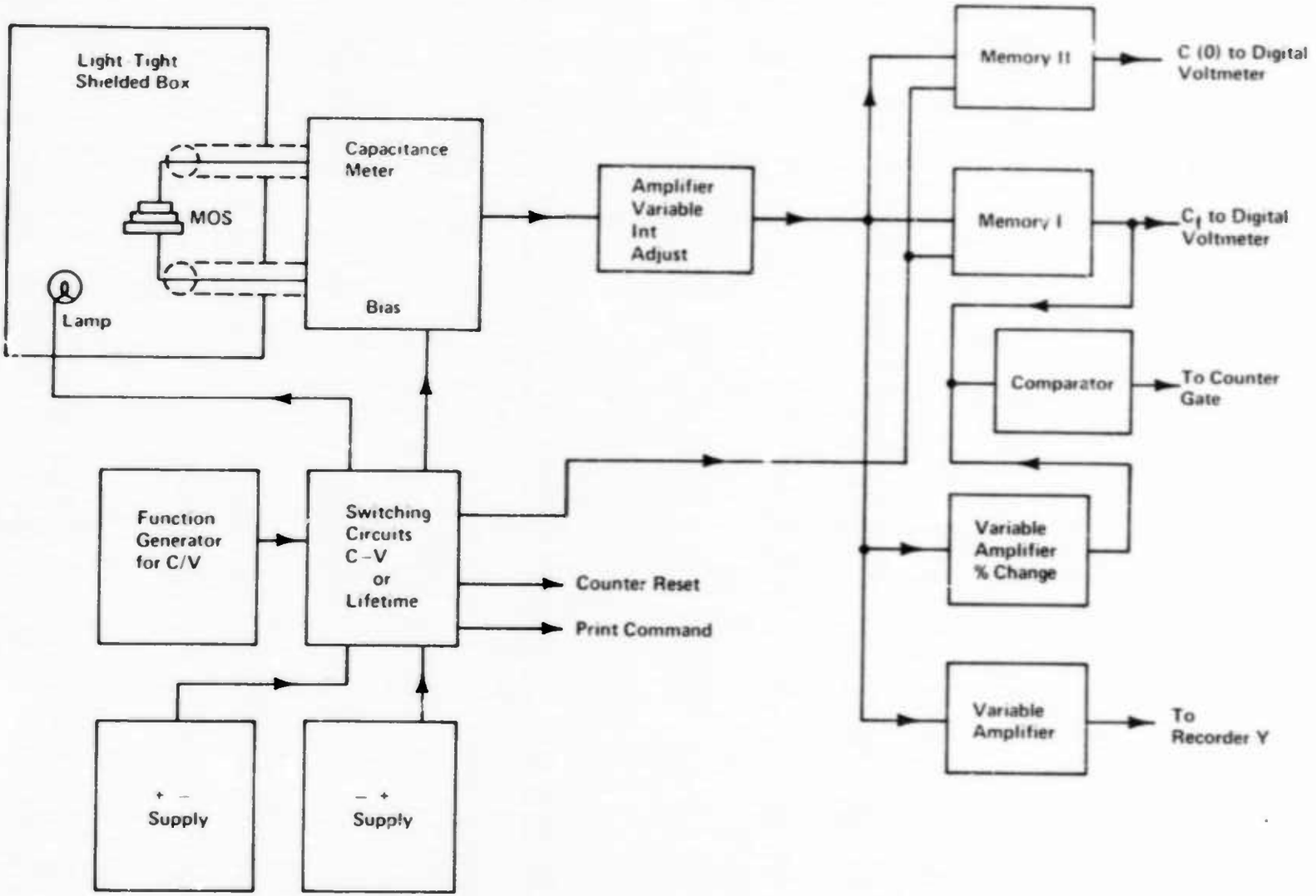


Fig. 6a. Block diagram of the "long" lifetime case.

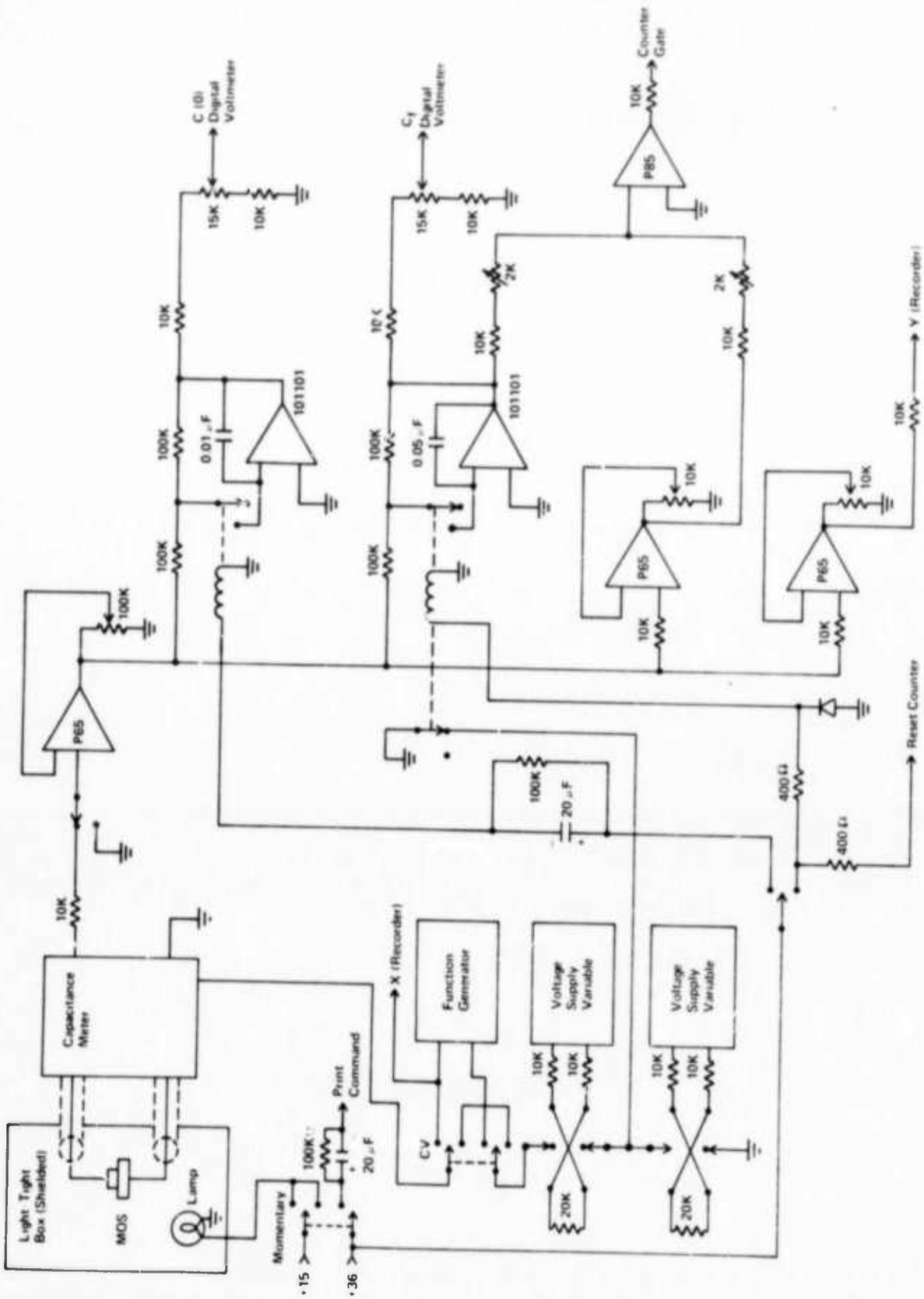


Fig. 6b. Schematic of the "long" lifetime case.

A similar setup is used for T-values > 1 msec (Fig. 7). The block diagram of this setup has been published by Princeton Applied Research, N.J.[7] In this case, the output of a lock-in is proportional to C_R. For $T \lesssim 1$ sec, the plotter must be replaced by an oscilloscope. We use a Tektronix sampling oscilloscope, which can be connected to a digital counter. We count the time between $t = 0$ and the instant $C_R = 0.8 \times (1 - C_R(0))$ (as an example).

The third setup is used for the "short" lifetime case $T < 4$ sec. As shown in Fig. 8, it consists of a series combination of the MOS capacitor, C, and a standard capacitor, $C_{st} >> C$. The voltage drop, V_{st}, across C_{st} is measured with an operational amplifier. The large parallel resistor, R, defines the d-c potential for the amplifier. The condition $R >> 1/\omega C_{st}$ yields the upper limit for the measurable T values: $T << R \cdot C_{st} \cdot 2\pi$. It is favorable to choose a large value of R rather than of C_{st}, because an increase in C_{st} implies a loss in resolution. We can write

$$V_{st} = V_a (1/C_{st}) / (1/C + 1/C_{st}),$$

$$V_{st} \approx V_a \cdot C/C_{st}. \qquad [11]$$

However, C is a mixture of the high- and low-frequency capacitance and relaxes finally to C_{ox}. Though it might be possible to calculate C (t) analytically, we prefer a different, more convenient, interpretation. We write the voltage drop, V_{MOS}, across C as

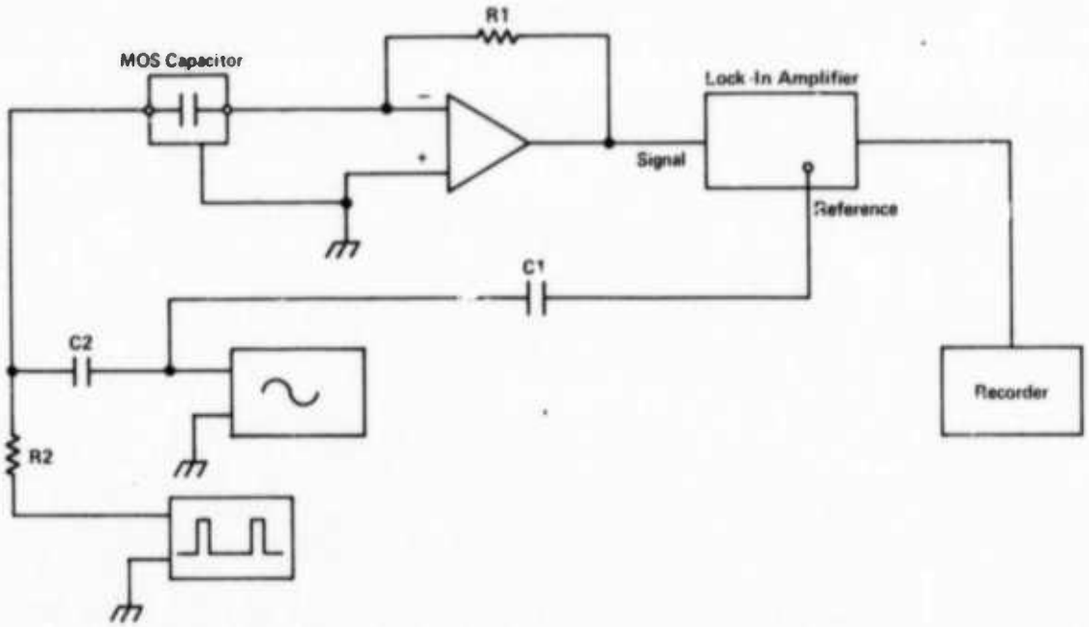


Fig. 7. Block diagram of the "medium" lifetime case.

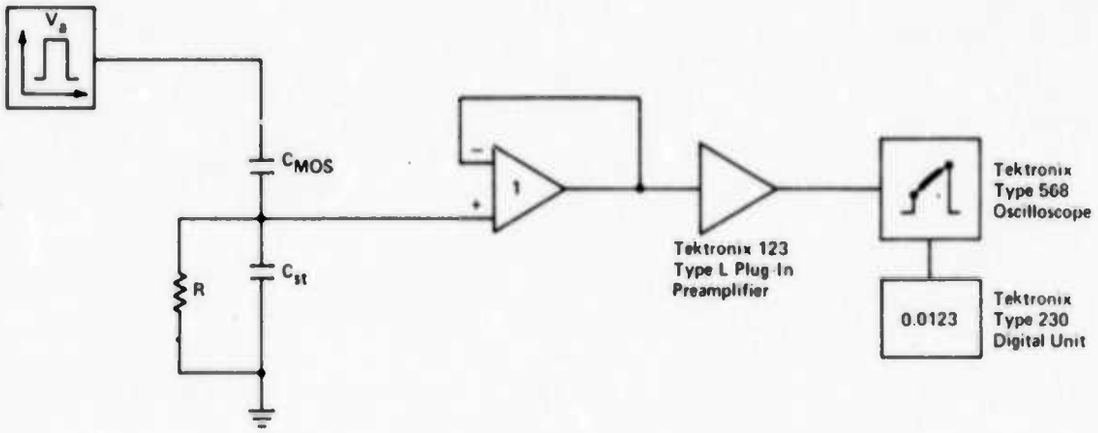


Fig. 8. Block diagram of the "short" lifetime case.

$$V_{MOS} = V_{ox} + \phi_s.$$

Thus,

$$V_a = V_{st} + V_{ox} + \phi_s.$$

Because C_{st} and C_{ox} are linear capacitors,

$$V_{st} \cdot C_{st} = V_{ox} \cdot C_{ox},$$

$$V_a = V_{st}(1 + C_{st}/C_{ox}) + \phi_s,$$

and, finally,

$$V_{st} = (V_a - \phi_s) / (1 + C_{st}/C_{ox}), \qquad [12a]$$

$$\phi_s = V_a - V_{st}(1 + C_{st}/C_{ox}). \qquad [12b]$$

By measuring V_{st}, we know ϕ_s and can use Eq. [8]. Note that, in Eqs. [11] and [12], V_a assumes two different values. For this reason, we do not obtain the same final values before and during the pulse as in Fig. 4, but two values as shown in Fig. 9. The time measurement is again performed with a digital counter connected to a sampling scope. The advantage in using the scope and counter is the speed and accuracy of the reading. The error in reading without the counter can be larger than 100%, especially for small ΔV_a signals. Figure 10 shows how the time marks of the counter are set. The left time mark is set at the onset of the pulse and indicates the start of the counting (left arrow).

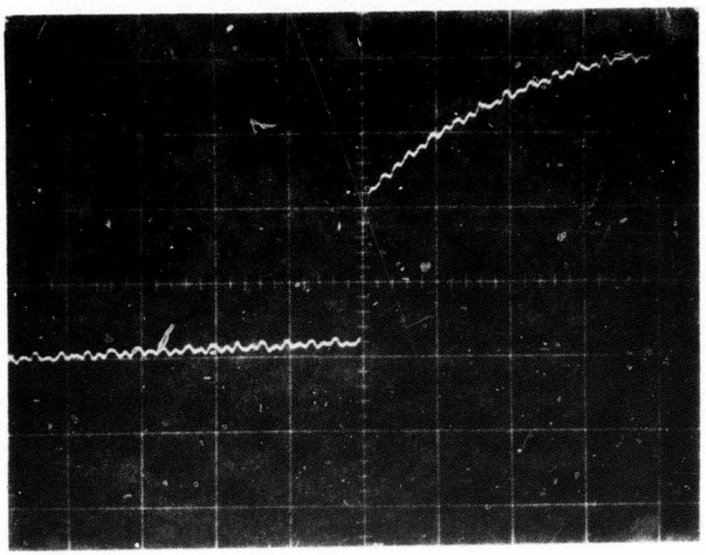


Fig. 9. Photograph of a C-t trace of a low-frequency measurement. Note that the scale factor V_a changes when the pulse is switched on (50 msec/div. horiz.).

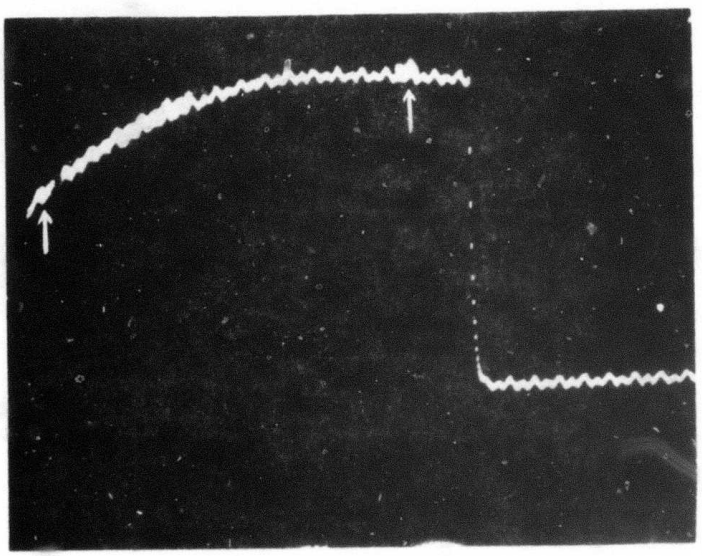


Fig. 10. Photograph of the same sample as in Fig. 9. The time marks and the portion of the curve during which the counter operates are intensified. The sequence of the pulse is reversed compared with Fig. 9.

The stop time mark can be seen at the end of the pulse (right arrow). The time of an 80% vertical transition between the zones is measured as indicated by the heavy line. Note that we are within the upper limit for the RC_{st} ($10^7 \Omega \times 10^{-7}$ F) combination, as shown by the slight tilt of the last horizontal line.

In Figs. 11 and 12, the results of two measurements are fitted to the thermal relaxation curve described by Eq. [10]. Good agreement is obtained in Fig. 11, whereas, for the long-lifetime silicon in Fig. 12, the agreement is merely satisfactory. This is in accord with the general observation that in this range ($\tau \approx 1$ msec) the error becomes 100%. A Zerbst plot (Fig. 13) taken from Fig. 11 gives $\tau = 2.5 \times 10^{-3}$ sec and $s_o = 3.7 \times 10^{-2}$ cm/sec. This technique yields $\tau = 1.3 \cdot 10^{-3}$ sec with $C_{ox} = 130$ pf, $C_f = 75$ pf, $C_R(0) = 0.339$, and $t = 1060$ sec for a selected level $C_R(t) = 0.8$.

These experiments were repeated with different samples. Following are typical results obtained:

τ (Zerbst) = 420 μsec - τ (this technique) = 252 μsec,
τ (") = 144 μsec - τ (" ") = 108 μsec,
τ (") = 5.6 μsec - τ (" ") = 4.6 μsec,
τ (") = 1 msec - τ (" ") = 850 μsec.

Furthermore, we examine how a nonthermal generation mechanism affects the reliability of this technique. This is done by shining controlled intensities of light on the sample or by using mechanically or chemically stressed substrates

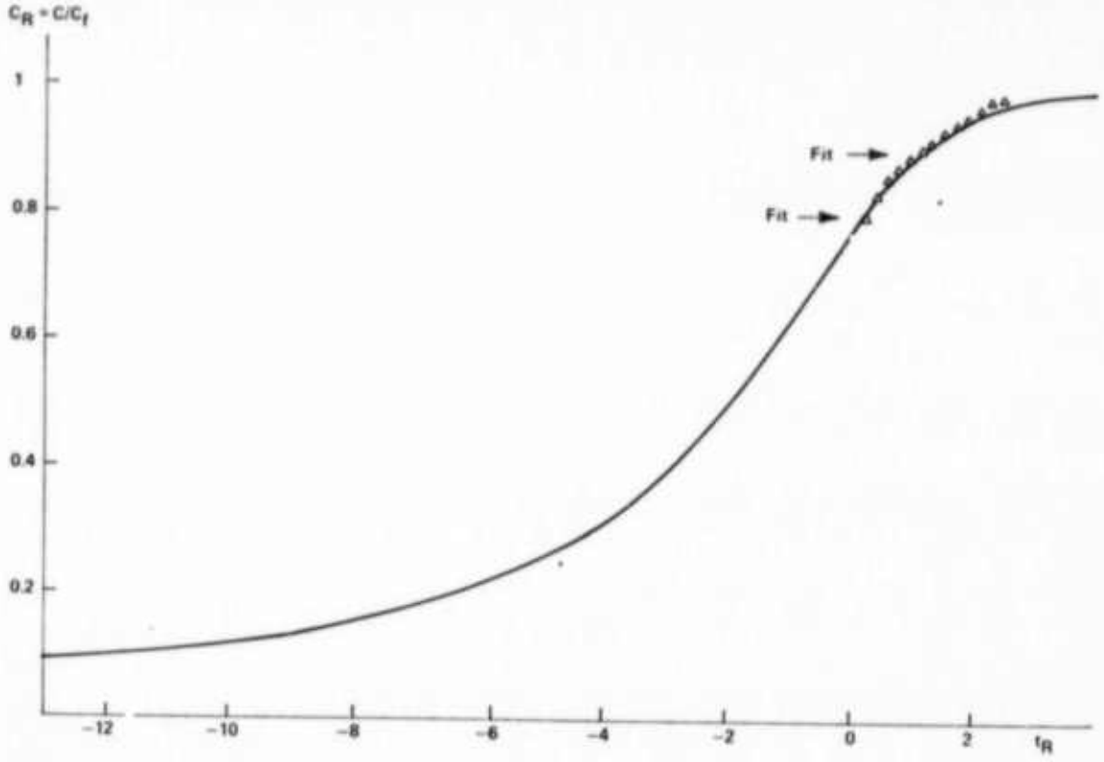


Fig. 11. C-t plot of a sample fitted to the thermal relaxation curve.

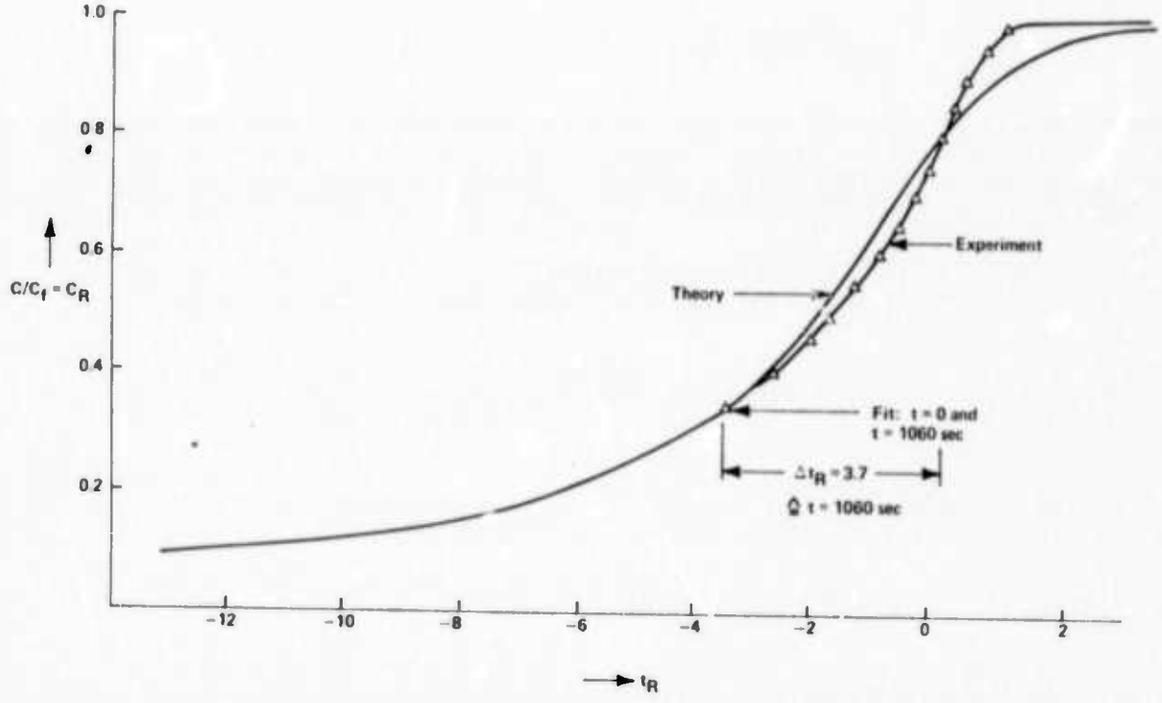


Fig. 12. Same curve as in Fig. 11, taken from another sample.

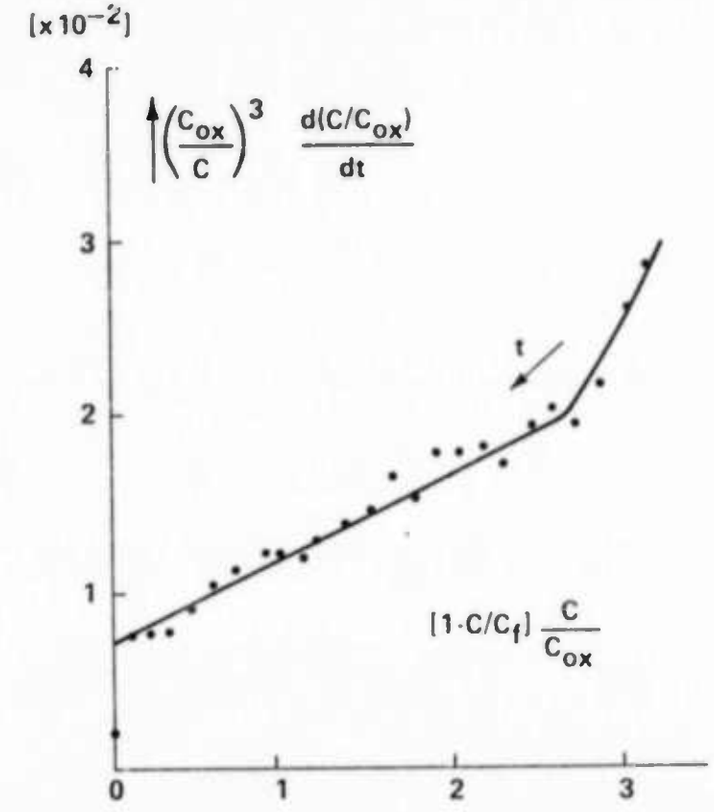


Fig. 13. Zerbst plot for the sample of Fig. 12.

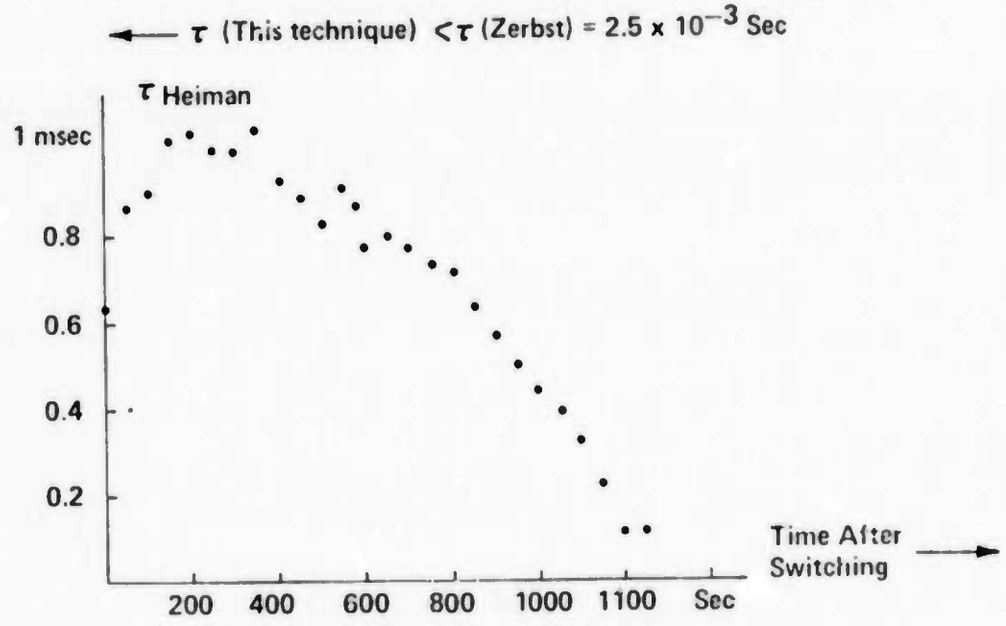


Fig. 14. Heiman plot for the sample of Fig. 12.

for the MOS fabrication. The general result is as follows. Whenever a reasonable Zerbst plot is obtained, a lifetime results which agrees with the value of our technique. However, when the Zerbst plot fails, our technique still reports at least an "effective" lifetime, which can be used for process monitoring.

In Fig. 14, we show an example of a Heiman plot. (We define a Heiman plot as the τ vs t curve, where τ has been obtained from a C-t plot with the procedure proposed by Heiman;[5] of Eq. [9].) The data are taken from the sample used in Fig. 12. It can be seen that the agreement is poor.

It is possible to make the measurements and the data evaluation even more expedient by calculating, rather than measuring, $C_R(0)$. For this purpose, we use the following model: a fast-increase dV_a of the applied voltage displaces a charge $dQ = C(V_a)x\ dV_a$ across the capacitor, C. Only the majority carriers can follow, and a space charge $q \cdot N_A \cdot dx_{Dl}$ matches dQ. Thus, C consists of the series combination C_{ox} and ϵ_{si}/x_{Dl} with the initial condition $x_{Dl} = x_{Do}$ or $C = C_f$ for $t = 0$. (Because of this initial condition, we cannot assume $\epsilon_{si}/x_D >> C_{ox}$ as done by others [8] in solving Eq. [7].) The solution of the differential equation

$$N_A \cdot q \cdot dx_{Dl} = (1/C_{ox} + x_{Dl}/\epsilon_{si})^{-1} \cdot d\,V_a$$

is

$$N_A \cdot q\, x_{Dl}^2/(2 \cdot \epsilon_{si}) + N_A\, q\, x_{Dl}/C_{ox} = \Delta V_a + N_A q x_{Do}^2/(2\epsilon_{si}) + N_A \cdot q \cdot x_{Do}/C_{ox}$$

or

$$x_{Dl} = (-\epsilon_{si}/C_{ox}) + \sqrt{2\,\Delta V_a \cdot \epsilon_{si}/(N_A \cdot q) + ((\epsilon_{si}/C_{ox}) + x_{Do})^2}\ . \qquad [13]$$

We obtain a depletion capacitance

$$C_D(0) = \epsilon_{si}/x_{Dl}$$

and

$$C_R(0) = (C_{ox} \cdot C_D(0)/(C_{ox} + C_D(0)))/C_f\ . \qquad [14]$$

We check this model by switching an MOS capacitance with ΔV_a values = 1, 2, 3...10 volts (V). The bias V_a is 20V and 40V. The results and the comparison with the theoretical values according to Eq. [14] can be seen in Fig. 15. As expected, there is no measurable difference between the 20V and 40V measurements. The systematical discrepancy between the experimental and theoretical values can be explained by the error in determining the doping concentration. Different measurements give a value between $8 \cdot 10^{14}$ and $1.3 \cdot 10^{15}/cm^3$. The same model can be used to calculate $\phi_R(0)$. It must be emphasized that this model ignores some effects that might reduce the value of $C_R(0)$. Among these, the lateral current paths and the finite thickness of the inversion layer, probably, are most important.

SUMMARY

A new fast technique for large-scale measurements of silicon lifetimes is described. The error limit, compared with a Zerbst plot, is less than 20% for $\tau \leq 10\ \mu$sec and increases to $\approx 100\%$ for $\tau \approx 1$ msec. This error limit can be reduced by choosing a smaller stopping level $C_R(t)$. For the long-lifetime case, the measurement time is determined by the setting of the selected level and by the τ value of the silicon itself. For the short-lifetime case, it is determined by the time required to find the optimum pulse frequency. The com' ater time can be disregarded.

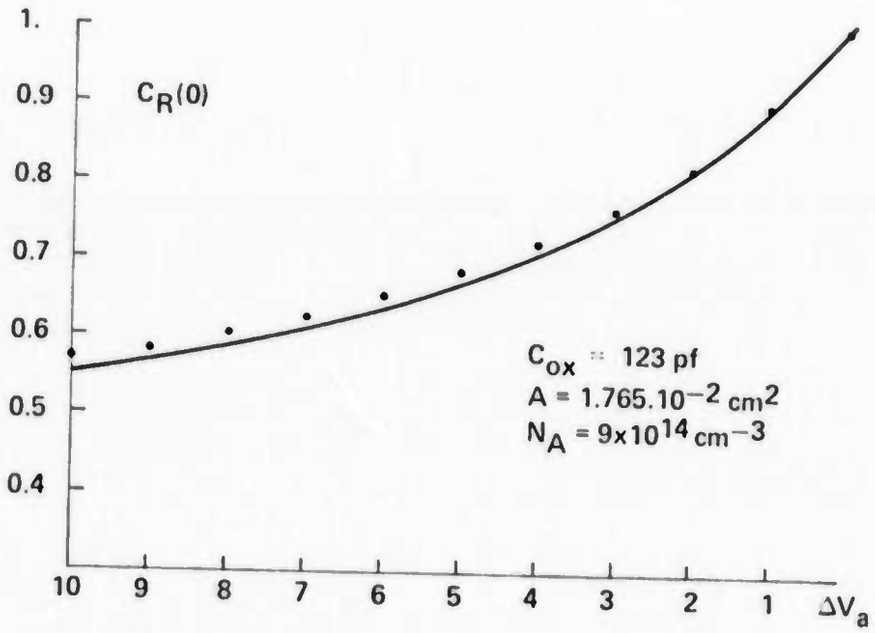


Fig. 15. $C_R(0)$ vs ΔV_a, experiment: theory: solid line. $C_{ox} = 123$ pf, $A = 1.765.10^{-2}$ cm^2, and $N_A = 9 \times 10^{14}$ cm^{-3}.

ACKNOWLEDGMENTS

The authors are grateful to Dr. G. H. Schwuttke for his continuous encouragement and support of this work.

A setup similar to our low-frequency setup has been developed independently by K. Ziegler and E. Klausmann for the measurement of equilibrium surface potentials. We are obliged to them for communicating their results in advance of publication.

REFERENCES

1. M. Zerbst, Z. Angew. Phys., 22, 1, 30 (1966).
2. P. Tomanek, Solid-State Electron., 12, 301 (1969).
3. S. R. Hofstein, IEEE Trans., ED-14, 785 (1967).
4. D. K. Schroder and H. C. Nathanson, Solid-State Electron., 13, 577 (1970).
5. F. P. Heiman, IEEE Trans., ED-14, 11, 781 (1967).
6. Y. Kano and A. Shibata, Jap. J. Appl. Phys., II, 8, 1161 (1972).
7. Princeton Appl. Res. Tech. Note TN102.
8. J. S. T. Huang, Proc. IEEE, 58, 11, 1489 (1970).

NOTATION

Symbols

C	measured capacitance
C_f	minimum capacitance of high-frequency C-V curve
C_D	depletion capacitance
C_{ox}	oxide capacitance
C_I	inversion capacitance
C_{st}	standard capacitance
ϕ_s	surface potential
V_a	applied voltage
V_{ox}	voltage drop across the oxide
Q_I	inversion charge
Q_D	depletion charge
N_A	doping density
ϵ_{si}	dielectric constant of silicon
q	elementary charge
x_D	depletion width
τ	lifetime
t	time
s_o	surface generation velocity
n_i	intrinsic carrier density

Subscripts

a	applied
ox	oxide
f	final, in equilibrium
th	thermal
Gen	generation
s	surface
R	in reduced units
D	depletion